GenePool: A Cloud-Based Platform for Interactive Visualization and Integrative Analysis of Genomics and Clinical Data

Hua Fan-Minogue, MD, PhD[1*], Marina Sirota, PhD[1*], Sandeep Sanga, PhD[2*], Dexter Hadley, MD, PhD[1], Atul J. Butte, MD, PhD[1], Tod Klingler, PhD[2¶]

[1]Division of System Medicine, Department of Pediatrics, Stanford School of Medicine, Stanford, CA; [2]Station X Inc, San Francisco, CA

Abstract

Advances in genomic technology harnessed by large-scale efforts from interdisciplinary consortia have generated more comprehensive genomic data and provided unprecedented opportunities for understanding human health and disease. The Cancer Genome Atlas (TCGA) in particular contains comprehensive clinical, genomics and transcriptomics measurements across over 16,000 of patient samples and over 30 tumor types. However, the vast volume and complexity of these data has also brought challenges to biomedical research community for translating them into insights about diseases quickly and reliably. Here we present a cloud-based platform, GenePool, which allows rapid interrogation of multi-genome data and their associated metadata from public or private datasets in a secure and interactive way. We demonstrate the functionality of GenePool with two use cases highlighting interactive sample browsing and selection, comparative analysis between disease subtypes and integrated analysis of genomic, proteomic and phenotypic data. These examples also display the promise of GenePool to accelerate the identification of genetic contributions to human disease and the translation of these findings into clinically actionable results.

Introduction

Since the completion of the Human Genome Project, tremendous progress in the technology has enabled the generation of genomic data in a remarkably high throughput and cost-effective manner, which is strikingly illustrated by the next-generation sequencing technologies (1). The resulting unprecedented level of sequence data has facilitated the identification of genes associated with disease or drug response, some of which have led to improved therapies (2). Large interdisciplinary consortia have played a major role in harnessing these advances and building comprehensive catalogues of genomic data. For example, The Cancer Genome Atlas (TCGA) characterizes cancer transcriptomics and genomics data along with phenotypic data (3), the 1000 Genome Project focuses in depth on the human genetic variation (www.1000genomes.org), and The Genotype-Tissue Expression (GTEx) program aims to identify the correlations between genetic variation and gene expression in multiple tissues (commonfund.nih.gov/GTEx). Various types of data, including clinical, genetic variation, mRNA expression, methylation, copy number variation, and others have been generated and released as resources to the community.

The rapid growth in generation of these sorts of large-scale genomics efforts brings new challenges to biomedical research. The raw and processed data often reside in the repository of each individual consortium, the access of which needs manual downloading through individual data portal. This also leaves users to concern about the compatible computing power and data storage capacities. Furthermore, analysis of these data requires specialized bioinformatics knowledge about the computing algorithms and statistical methods appropriate to genomic data, which is not readily available in most biomedical research labs (4). Therefore, data management, query and analysis become new challenges for efficiently translating the abundance of data into knowledge about disease.

Some solutions are available that enable easy access to the data. For example, the data portal for the International Cancer Genome Consortium (ICGC) allows quickly browsing or query of genomics data from cancer projects, including TCGA (5). COSMIC (catalogue of somatic mutations in cancer) (6) and cBioPortal (7), provides easy access to curated comprehensive information on cancer genomics data. However, these data portals have no or limited capability for empowering real-time data analysis. The Broad Institute (www.broadinstitute.org) offers a suite of online analysis tools for large genome-related datasets. Although these tools can be used against the existing datasets of Broad Institute, application on alternative datasets requires local downloading of the tools of choice. Cloud computing technology has been increasingly used to meet the needs for easy access to analysis tools, as well as storing and retrieval of large datasets. One example is the CloudMap (8), however, it is only applied to a single type of data and integrative analysis is not available. Therefore, a user-friendly solution for accessing, processing, analyzing and integrating genomic, clinical as well as other types of data is currently lacking.

Here we present a cloud-based platform, GenePool™ (https://stationx.mygenepool.com), which aims to provide a one-stop solution for genomic data analysis and integration. It offers secure storage, genomics workflows, interactive visualization and interpretation

[*] These individuals contributed equally

[¶] Corresponding author

of genomics data obtained from various genomic technologies, as well as associated metadata. It currently holds over 10 public datasets across several modalities, including TCGA and GTEx, and multiple biological annotation engines, such as Gene Ontology (GO) and Disease Ontology (DO). GenePool allows import of external genomic data for customized use. It also allows reporting and securely sharing of results with multiple options for annotation, sorting and data export. To demonstrate its function and usability, we applied GenePool to the TCGA datasets of the two subtypes of lung cancer, Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). For each subtype, we first performed interactive browsing and query, then carried out analysis of differential expression for each subtype. We identified genes that were unique to each subtype as well as a common lung cancer signature. In the second use case, we dove into a deeper analysis of a specific subgroup of LUAD patients focusing on individuals with a history of smoking across multiple modalities including gene expression, somatic mutation and protein expression data.

Methods

System architecture

GenePool runs on Amazon Web Services (http://aws.amazon.com/) and takes advantage of EC2 compute, S3 storage, and other AWS services to maintain a secure computing and storage environment (Figure 1a). Its backend code is primarily implemented in Java, which interacts with a statistical engine powered by R to run select computations. Its front-end primarily uses Javascript and HTML 5.0. It leverages a hybrid database layer to manage and store genomics and annotation data. This hybrid data layer consists of a PostgreSQL database and a flat-file, NoSQL-like component. The supported data types include expression data (RNAseq and miRNAseq) in BAM format or as tab-delimited files, and sequence variation (whole genome and exomes, copy number and DNA mehtylation) mutation calls in VCF format. Users are allowed to import genomics data and/or work with the reference library of GenePool. It has canned workflows, but also allows customized workflows developed with its APIs. The software solution has fully integrated multi-genome browser and provides an easy way to store and select genomics datasets according to sample-associated metadata for analysis, and an automated system for performing routine, well-characterized genomics workflows through an intuitive interface.

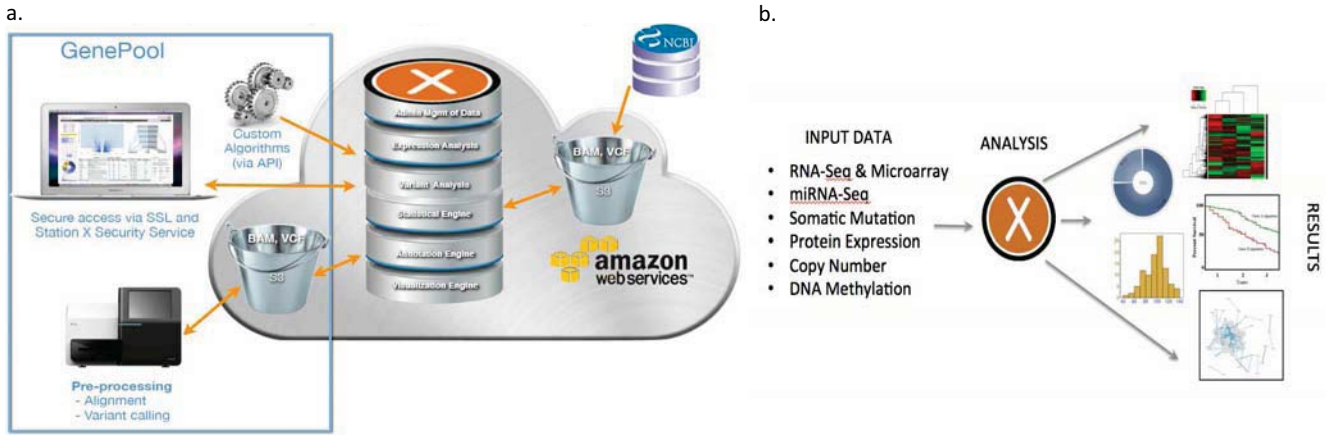


Figure 1. Overview of GenePool platform. a. System architecture, b. functionality.

Analysis tools

GenePool has well-characterized workflows for various types of genomic data analysis, such as expression profiling and comparisons, sequence variation profiling and comparisons, time-to-endpoint analysis, paired and trio analysis, gene enrichment analysis. A variety of statistical methods are available as options for the user to choose from, such as t-test or likelihood ratio test for expression comparisons; chi-squared test for variation comparisons; fisher's exact test for gene enrichment analysis; univariate cox proportional hazard ratio test for time-to-endpoint analysis; and q-values for estimating FDR. The results can be visualized in heat map, co-expression networks, kaplan-meier survival curves, scatter plots, histograms and bar charts (Figure 1b).

Scalability

The combination of Amazon Cloud Infrastructure and GenePool's own unique platform architecture grants it elastic scalability, where the accessibility can be scaled up or down depending on demand. Utilizing load-balancing and server / application metrics, GenePool can sense periods of high demand and add additional compute as required to the middle application tier. This additional compute is leveraged to meet the demands of customer workloads particularly during high concurrency and during period where large analyses are being performed. This capability allows GenePool to meet the ever changing and unpredictable demands of our users' workloads. This elastic scalability also benefits the data import process. As data import jobs can vary in size and computational requirements, each job is assessed for their computational needs and a unique server instantiated for the specific job.

This allows GenePool to vertically scale, applying large multi processor servers for large jobs so that data is imported quickly and efficiently. Depending on the type of data being imported, GenePool can load 250-500 samples per hour. In the future, GenePool will have the capability to spread import jobs across multiple servers (horizontal scalability) to achieve an even greater throughput for our customers with even larger data volumes and data processing requirements.

Results

In this paper we present two use cases of the GenePool platform in the context of analyzing LUAD and LUSC data from TCGA. The first use case focuses on integrating RNA-Seq expression data across different types of lung cancer, identifying gene expression signatures of LUAD and LUSC individually as well as a common signature for lung cancer. The second use case focuses on a more specific subpopulation of lung adenocarcinoma patients who smoked 60 or more packs per year vs. life-long non-smokers and integrates data across several different modalities including RNASeq, somatic mutation and protein expression.

Use cases

1. Interactive browsing and differential expression query of LUAD and LUSC TCGA data

GenePool allows intuitive, user-friendly sample selection based on patient and sample metadata as illustrated in Figure 2a. The samples and the corresponding meta-data are displayed in a table allowing the user to select or filter the dataset based on variables and annotations of interest. Pie charts and histogram widgets on the left side of the screen allow the user to sub-select populations of interest for further analysis. For instance one can quickly compute differential expression between the LUAD primary tumor and non-tumor samples (Figure 2b) and visualize the expression levels for a certain gene of interest (Figure 2c). We carried out differential expression between tumor and non-tumor samples for LUAD and identified 2379 genes that are significantly differentially expressed (abs(FC) > 3, FDR < 5.04e-07). We have carried out a similar analysis for LUSC identifying 4025 genes that are significantly differentially expressed (FC > 3, FDR < 5.04e-07). We identify specific subset of 1943 genes that are commonly significantly up- or down- regulated in both LUSC and LUAD such as subtype specific signatures for each subtype. KRT6A, a protein encoded by this gene is a member of the keratin gene family and has previously been associated with non-small cell lung cancer (9), is an example of a gene that is up-regulated in LUSC, but not LUAD (Figure 2c). SPINK1, serine peptidase inhibitor, Kazal type 1 is up-regulated in LUAD but not LUSC as recently reported by Lazar et al (10). Finally several known tumor markers such as PRAME, COL11A1 and various MMP genes are observed to be up-regulated in both tumor subtypes.

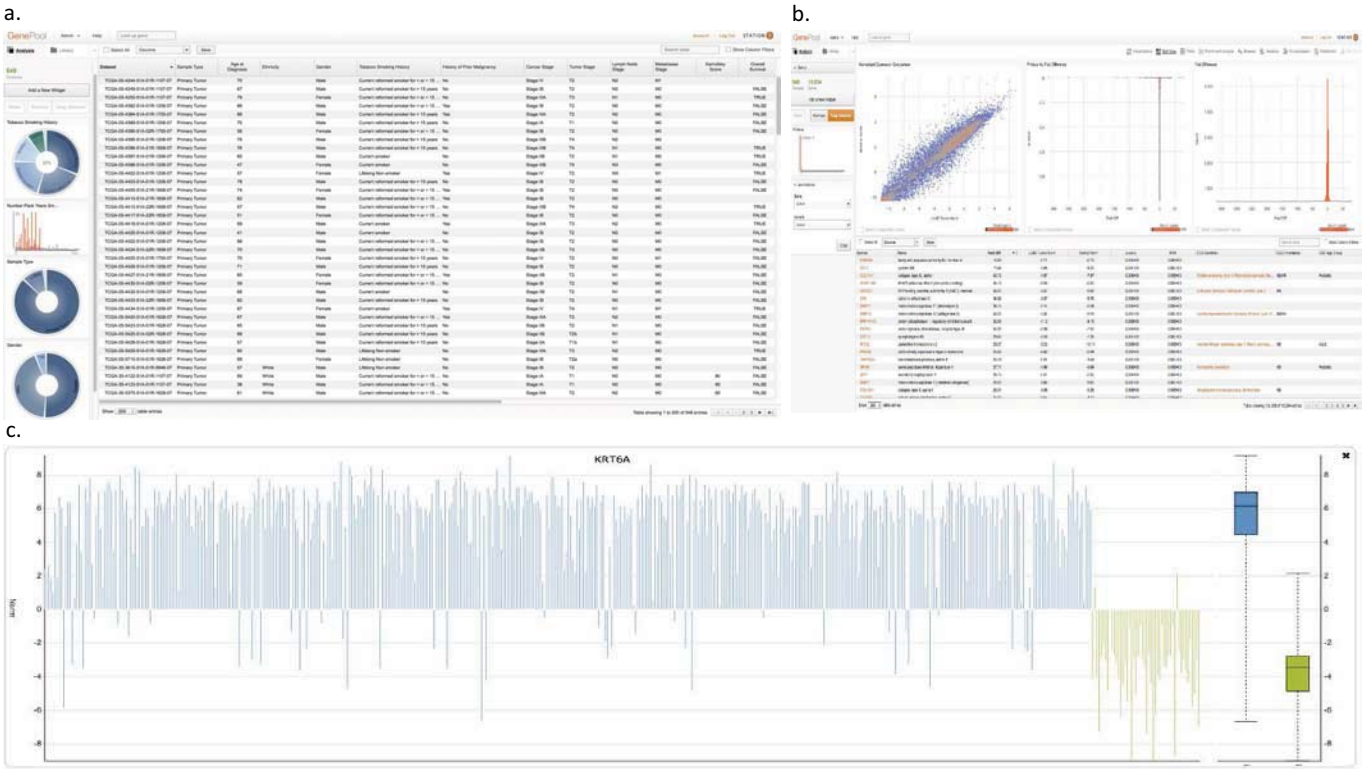


Figure 2. Using GenePool for sample selection and differential expression analysis. a. Sample selection based on meta-data. b. Differential expression analysis in LUAD sorted by fold change. c. Gene expression levels for KRT6A, which is up-regulated in LUAD (blue) in comparison to adjacent normal samples (green).

2. Characterizing sub population of LUAD patients across different data types

We have used GenePool to carry out more complex queries focusing on the analysis of LUAD subset of patients who smoked 60+ pack years in comparison to life-long non-smokers. We started by running a differential gene expression analysis with available RNA-Seq samples for both groups, specifically 64 smokers and 70 non-smokers. We then selected for the top differentially expressed genes based on statistical significance and fold change using the interactive tools, resulting in a set of 2,198 statistically significant genes (FDR ≤ 15%). Enrichment analysis of these top genes identifies "Cell Cycle"- related genes amongst other gene sets to be significantly enriched, an observation previously noted by other investigators (11, 12). We overlaid the gene ontology annotation with the gene expression to focus on the cell cycle-related genes and visualized those with a co-expression network (Figure 3a) confirming the similar expression profiles amongst cell cycle-related genes. We used the interactive heatmap feature to mine for sample and gene clusters, as well as overlay patient and sample metadata. In Figure 3b we confirm that the top differentially expressed genes cluster the samples according to patient smoking history. We then focus on the same subset of samples and look for biomarkers in smoking-related lung adenocarcinoma that are most associated with survival. By running a time-to-endpoint analysis in GenePool, we identify VEGFC, ERBB3, and other genes whose expression is statistically associated with survival. While ERBB3 has been previously associated with lung cancer survival (13), VEGFC's correlation with survival was previously not shown to be statistically significant (14). Furthermore, we carried out somatic mutation comparison in the same subset of samples (*i.e.,* smokers vs. non-smokers). The initial profiling of the two groups returned 40,277 unique somatic mutations. We used GenePool's interactive visualizations and tables to select for rare mutations (≤5% allele Frequency in the 1000 Genomes Project and Exome Sequencing Project) and have a functional impact on Cancer Gene Census genes maintained in the COSMIC database. In the Gene-level view, we see the mutational frequency of a few key genes with known associations to lung cancer: TP53 (71% in Smokers and 33% of Non-Smokers), KRAS (33% in Smokers, 16% in Non-Smokers), and EGFR (10% in Smokers, 33% in Non-Smokers) (15). Finally we compared protein expression between the smoker and non-smoker groups and found 26 out of 160 proteins profiled to be statistically differentiated (FDR ≤ 15%). We confirm ERBB3 is over-expressed at the protein-level in smokers(16), but interestingly was not statistically differentially expressed by the gene expression analysis.

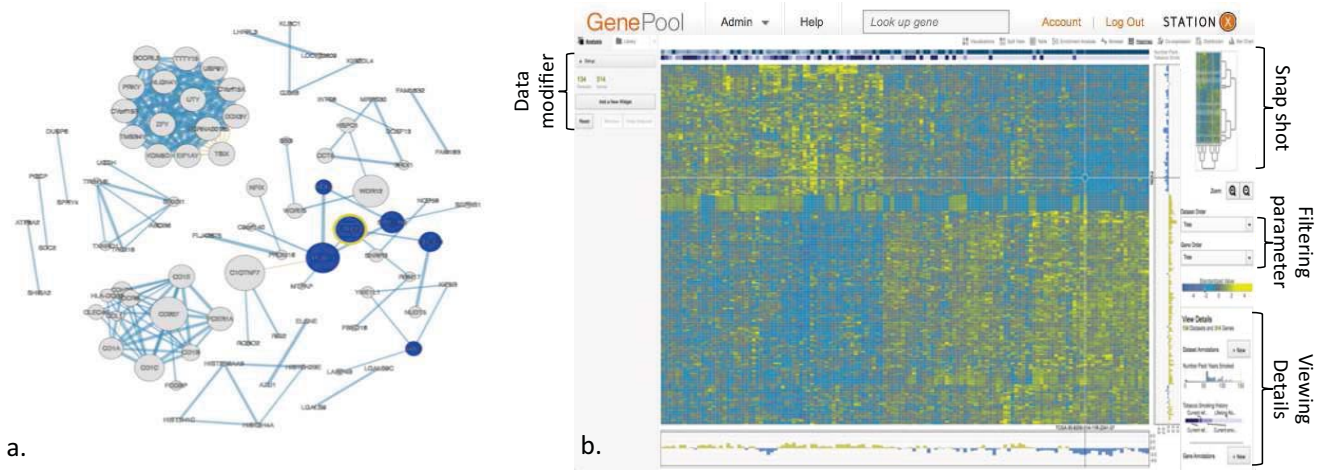


Figure 3. Characterizing sub population of LUAD patients across different data types. a. A co-expression network with cell-cycle specific genes highlighted in blue showing consistency of expression in cell-cycle related genes. b. Interactive heatmap (genes are on the y-axis and patients on the x-axis) showing that clustering based on the top differentially expressed genes results in samples clustering based on smoking history.

Conclusion

We have demonstrated GenePool, a cloud-based platform that greatly simplifies the analysis of the ever-increasing amount of human genomic data in the context of analyzing lung cancer data obtained from TCGA. We present the platform in the context of two use cases, which allow users to select out subsets of samples based on certain meta-data, carry out statistical analyses across different cancer types as well as data modalities. In particular we used GenePool to investigate and identify subtype-specific gene expression signatures for LUSC and LUAD as well as a common lung cancer signature to both subtypes. We also used GenePool to carry out integrative analysis of a subset of LUAD patients with a history of smoking in comparison to the non-smoker population looking for markers in differential gene and protein expression as well as somatic mutations patterns. In both cases we were able to carry out the analysis rapidly and identified both known and novel genomic markers that are of interest to the research community. This platform, thus, provides a comprehensive solution to the scientific community for rapidly utilizing public data resources of patient-derived genomic data. In the future, GenePool will keep updating its reference library and diversifying its workflows, with the goal to accelerate the translation of genomics potential into actionable clinical solutions.

Acknowledgements

Partial support for this research was provided by the Lucile Packard Foundation for Children's Health and the Stanford Child Health Research Institute. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under award number R01 CA138256. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. MacArthur DG, Lek M. The uncertain road towards genomic medicine. Trends in genetics : TIG. 2012;28(7):303-5.
2. Green ED, Guyer MS, National Human Genome Research I. Charting a course for genomic medicine from base pairs to bedside. Nature. 2011;470(7333):204-13.
3. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.
4. O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. Journal of biomedical informatics. 2013;46(5):774-81.
5. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database : the journal of biological databases and curation. 2011;2011:bar026.
6. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research. 2011;39(Database issue):D945-50.
7. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery. 2012;2(5):401-4.
8. Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O. CloudMap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics. 2012;192(4):1249-69.
9. Camilo R, Capelozzi VL, Siqueira SA, Del Carlo Bernardi F. Expression of p63, keratin 5/6, keratin 7, and surfactant-A in non-small cell lung carcinomas. Human pathology. 2006;37(5):542-6.
10. Lazar V, Suo C, Orear C, van den Oord J, Balogh Z, Guegan J, et al. Integrated molecular portrait of non-small cell lung cancers. BMC medical genomics. 2013;6:53.
11. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, et al. Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PloS one. 2008;3(2):e1651.
12. Wu C, Zhu J, Zhang X. Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma. BMC bioinformatics. 2013;14:365.
13. Aurisicchio L, Marra E, Roscilli G, Mancini R, Ciliberto G. The promise of anti-ErbB3 monoclonals as new cancer therapeutics. Oncotarget. 2012;3(8):744-58.
14. Zhan P, Wang J, Lv XJ, Wang Q, Qiu LX, Lin XQ, et al. Prognostic value of vascular endothelial growth factor expression in patients with lung cancer: a systematic review with meta-analysis. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer. 2009;4(9):1094-103.
15. Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell. 2012;150(6):1121-34.
16. O'Donnell RA, Richter A, Ward J, Angco G, Mehta A, Rousseau K, et al. Expression of ErbB receptors and mucins in the airways of long term current smokers. Thorax. 2004;59(12):1032-40.

A prototype software pipeline to identify mutated genes that have a similar effect on tumor transcription

Stephen R. Piccolo

Department of Biology, Brigham Young University, Provo, UT (USA)

Abstract

Genome-wide studies have shown that a wide array of somatic mutations occur in non-small cell lung cancers. These mutations influence tumor growth via altering signaling cascades. Patient responses to treatments are highly variable, perhaps because different cascades are affected or because different components within a given cascade have different downstream effects. This manuscript describes a software pipeline for parsing somatic mutation data and grouping mutations according to similarities in gene expression for samples that either carry or do not carry mutations in a given gene. Genes that show relatively high similarity in gene expression are considered to have similar effects on tumor biology and thus may respond similarly to treatments. The tool's utility is illustrated via examining the effects of KRAS and EGFR mutations on gene expression, using tumor data from The Cancer Genome Atlas.

Introduction

Lung cancers traditionally have been classified according to the histology of tumor cells, as observed via microscopy. This approach has helped to define two main classes: small-cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC)(1). Within the latter group, three main subtypes have been identified. The two most prevalent of these subtypes are adenocarcinoma and squamous-cell carcinoma(2). A large proportion of both subtypes occur in people who have been tobacco smokers; however, many cases also occur in never-smokers(3). Mutations in various genes have been implicated in lung tumorigenesis. For lung adenocarcinomas, these genes include TP53, KRAS, EGFR, and STK11(4). The same genes are also mutated in lung squamous cell tumors(5); however, mutation frequencies differ considerably between these tumor subtypes(4,5).

An important challenge in evaluating such tumors is to group them according to how well each tumor might respond to specific therapies. Due to advances in DNA sequencing, it has become possible to profile tumors in a genome-wide fashion, in search of somatic mutations that influence not only tumor growth but also treatment responses. Hundreds of tumors have been examined, and key genes have been identified. In many cases, these mutations occur in a mutually exclusive manner within a given biological pathway such that a given tumor only accumulates a single mutation among the genes that drive a particular biological function. Thus in many cases, it is not enough to focus on any particular gene to understand what drives tumorigenesis and treatment responses. Instead it is helpful to look at many genes known to interact within a pathway context. Many "canonical" cancer pathways have been characterized via biochemical analysis; one example is the "Ras/Raf/MEK/ERK"

pathway, which governs cellular proliferation, differentiation, and survival(6). Mutations have been identified in various types of tumor tissues, including lung, for the proteins in this cascade(7). In addition, mutations commonly occur in EGFR, which is upstream of this pathway. However, although genes within a given pathway may interact with each other, it is likely that mutations in different pathway genes cause different downstream effects, in part due to crosstalk among pathways(8).

Many studies are being performed to understand how specific somatic mutations affect biomedical outcomes. However, due to various types of confounding effects, it may be difficult to make such connections reliably. One approach that may help draw closer to that goal is to evaluate the impacts of somatic mutations on RNA expression levels, which act as an intermediary. Somatic mutations may influence cellular behavior when they alter transcript sequences, splicing patterns, or expression levels. By examining such effects across many genes, it may be possible to identify expression patterns that result from such mutations and then to identify genes that have similar effects when mutated.

I developed a prototype software pipeline that characterizes the downstream effects of somatic mutations in tumors. To demonstrate how this tool might be useful in informing biomedical evaluations, I examined relationships between genes in the Ras/Raf/MEK/ERK pathway, other genes that interact with this pathway, and additional genes that are not believed to play a direct role in this pathway. For simplicity, this demonstration focuses primarily on lung adenocarcinomas and squamous-cell carcinomas; however, the pipeline can be applied to other tumor types and to other diseases for which such data are available.

Methods

Raw RNA-Sequencing data from The Cancer Genome Atlas (TCGA) were downloaded via the Cancer Genomics Hub (https://cghub.ucsc.edu) for 978 tumor samples. This included 489 samples for lung adenocarcinoma and 489 samples for lung squamous-cell carcinoma. The data were aligned and summarized using the *Rsubread* package(10), and RPKM values were calculated for each gene using the *edgeR* package(11). Because the analyses in this study examined RNA expression effects resulting from activation of the Ras superfamily, the RNA-Sequencing data were limited to 261 genes that have been shown to exhibit differential expression after RAS activation(12).

Somatic mutation data were downloaded from https://www.synapse.org/#!Synapse:syn1729383; following the instructions at https://www.biostars.org/p/91806, the mutations were filtered and annotated(13). Any mutation that the *snpEff* tool(14) indicated as having a low likelihood to impact protein sequence was excluded. Mutations that both SIFT(15) and Polyphen2(16) suggested to have a benign effect on protein function were also excluded.

The pipeline parses the somatic mutation data and identifies tumor samples that have a mutation in a specified gene. RNA expression values are then compared between samples that have a mutation in the gene and samples that have a mutation in another specified gene. A Euclidean distance measure is used to quantify the similarity in expression between the groups. Gene pairs for which median RNA expression distances are relatively small are considered to have similar downstream effects. Differences in distance were compared using an analysis of variance test.

The prototype is implemented as a *bash* script, which invokes *Python* and *R* code. All source code and scripts that were used for the analysis can be accessed at https://bitbucket.org/srp33/mutationexpressionprototype. The pipeline is designed to run on UNIX-based operating systems. Python version 2.7.5 and R version 3.1.1 were used for the analyses described in this paper.

Results

Of 171 lung adenocarcinoma samples for which mutation data were available, 46 (26.9%) had mutations in the KRAS gene. The protein associated with this gene interacts with genes in the Raf kinase family and can be activated indirectly via EGFR activation. Accordingly, I hypothesized that tumors with KRAS mutations would have RNA expression patterns similar to tumors with mutations in either EGFR or B-RAF. I also compared against samples that had a mutation in TP53 or STK11, which both are mutated relatively frequently in lung tumors but are not known to interact directly with the Ras/Raf/MEK/ERK cascade. Samples with mutations in KRAS showed the highest similarity to samples with mutations in EGFR (see Figure 1A). In a follow-up comparison (see Figure 1B), EGFR showed relatively high similarity to ERBB2 and ERBB4---which are homologs from the same protein family. These findings coincide with previous observations that mutations in KRAS and EGFR are usually mutually exclusive. This suggests that tumors with mutations in KRAS may respond to anti-EGFR treatments; however, research to date has been inconclusive on this topic(17). Perhaps surprisingly, samples with mutations in B-RAF had relatively dissimilar expression to KRAS-mutated samples, even though Ras proteins interact directly with the Raf kinase family.

Next I combined data from lung adenocarcinomas and squamous-cell carcinomas. Again, expression levels for KRAS-mutated samples were relatively similar to samples with EGFR mutations (see Figure 1C). However, the median distance between KRAS and TP53 mutated samples was considerably larger than for adenocarcinomas alone. TP53 mutations are highly prevalent in squamous-cell tumors(5), whereas KRAS mutations are much more prevalent in adenocarcinomas. This finding indicates that these mutations may cause systematic differences in tumor biology between these cancer types. However, further study is needed to examine the effects of batch variation and tissue-specific signals on such findings.

Discussion

A popular practice among genomics researchers is to define cancer subtypes based on gene-expression profiles. This approach has gained much traction among researchers studying breast and brain tumors(18,19). However, in many cases, somatic mutations are often the underlying drivers of altered cellular behavior and thus should be considered in determining cancer subtypes. Due to genetic heterogeneity and relatively low frequencies of many mutations, it may be helpful to group mutations that have similar downstream effects. This pipeline uses gene-expression profiles to represent such effects.

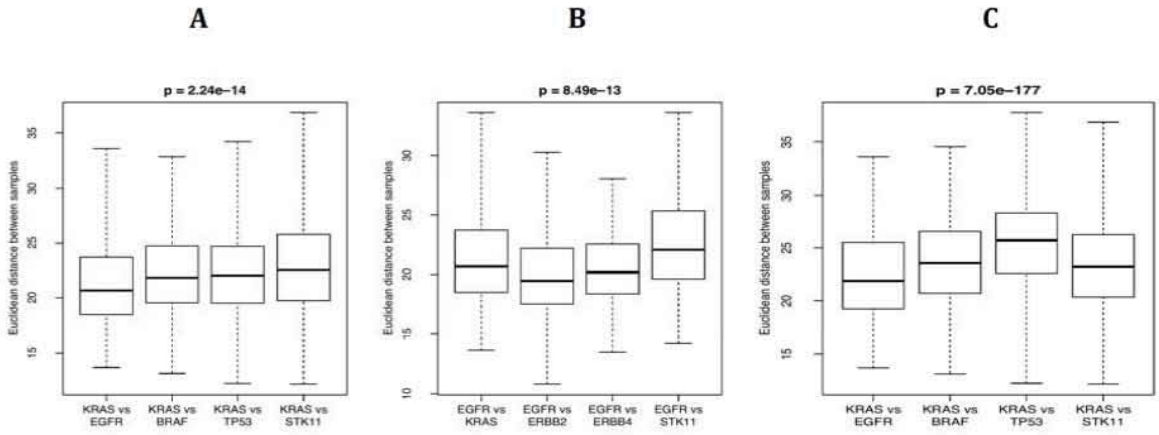


Figure 1: Euclidean distances between lung adenocarcinoma samples (A & B) that carried a somatic mutation in KRAS and samples that carried a mutation in various other genes. C shows data for lung adenocarcinoma samples combined with lung squamous-cell samples.

This pipeline could be applied in various additional ways. For example, it could group patients by mutations that occur at specific loci within genes rather than consider all mutations within a gene to have a similar effect. In addition, it could account for other types of (epi)genomic aberrations that may influence RNA expression, including copy-number variations, structural rearrangements, microRNAs, and DNA methylation changes. Because the tool uses input formats that are not specific to any type of genomic data, these other types of data could easily be substituted. Findings from this tool could also be combined with clinical data (much is provided in TCGA) to generate hypotheses about how these factors influence treatment responses, survival times, etc. Lastly, the choice to use the Euclidean distance measure was arbitrary; other, more sophisticated distance measures may be less sensitive to noise that inevitably affects RNA expression data.

References

1. Kumar V, AK A, JC A. Robbins Basic Pathology. 9th ed. Elsevier Saunders; 2013.

2. Lu C, Onn A, Vaporciyan A. Holland-Frei Cancer Medicine. 8th ed. Kong WK, Bast Jr. RC, Hait WN, Kufe DW, Pollock RE, Weichselbaum RR, et al., editors. People's Medical Publishing House; 2010.

3. Subramanian J, Govindan R. Lung cancer in never smokers: a review. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2007 Feb;25(5):561–70.

4. Network TCGAR. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014 Jul;511(7511):543–50.

5. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012 Sep;489(7417):519–25.

6. Kolch W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. The Biochemical journal. 2000 Oct;351 Pt 2:289–305.

7. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). Current Protocols in Human Genetics. Wellcome Trust Public Access; 2008;Chapter 10:Unit 10.11.

8. Osborne CK, Shou J, Massarweh S, Schiff R. Crosstalk between estrogen receptor and growth factor receptor pathways as a cause for endocrine therapy resistance in breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2005 Jan;11(2 Pt 2):865s–s.

9. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. Nature reviews Drug discovery. 2005 Dec;4(12):988–1004.

10. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic acids research. 2013 May;41(10):e108.

11. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England). 2010 Jan;26(1):139–40.

12. Bild AH, Chang JT, Joshi M-BB, Yao G, Lancaster JM, Wang Q, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. Nature Publishing Group; 2006 Jan;439(7074):353–7.

13. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013 Oct;502(7471):333–9.

14. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92.

15. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols. 2009 Jan;4(7):1073–81.

16. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics / editorial board, Jonathan L Haines [et al]. 2013 Jan;Chapter 7:Unit7.20.

17. Roberts PJ, Stinchcombe TE. KRAS mutation: should we test for it, and does it matter? Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2013 Mar;31(8):1112–21.

18. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2009/02/11. 2009 Mar;27(8):1160–7.

19. Lee S, Piccolo SR, Allen-Brady K. Robust meta-analysis shows that glioma transcriptional subtyping complements traditional approaches. Cellular oncology (Dordrecht). 2014 Aug;

Computational Phenotyping from Electronic Health Records across National Networks

Joshua C. Denny, MD, MS[a]; Abel Kho MD, MS[b]; Jyotishman Pathak, PhD[c]; Jimeng Sun, PhD[d]; Rachel L. Richesson, PhD[e]

[a]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA
[b]Departments of Medicine and Preventive Medicine, Northwestern University, Chicago, IL. USA
[c]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[d]School of Computational Science and Engineering, Georgia Institute of Technology, GA, USA
[e] School of Nursing, Duke University, Durham, NC, USA

Abstract

Wide-spread adoption of electronic health records (EHRs) containing rich amounts of longitudinal clinical data has led to expanded opportunities to repurpose these data for clinical and genomic research. We will present experiences and lessons learned across five national networks, including the Electronic Medical Records & Genomics (eMERGE) Network, Strategic Health IT Advanced Research Project (SHARPn), Pharmacogenomics Research Network (PGRN), Pharmacogenomics in Very Large Populations (PGPop), and PCORnet. We will also discuss the future of phenotyping research and its impact to both clinical research and operation. Topics covered will include strengths and weaknesses of EHR data for secondary research; computational approaches to phenotype representation and applications; novel methods of creating computational phenotypes; refining and sharing phenotypes; phenotype validation; and the standardization, interoperability, access, management, and governance of phenotype definitions across clinical sites and research networks. We will also discuss informatics successes and clinical and genomic discoveries, as well as current challenges using case studies from aforementioned networks.

Keywords: electronic health records, phenotyping, genomics, natural language processing, clinical research networks

Introduction

The adoption of EHRs has been one of the most important technological advances in healthcare. EHRs represent a robust source of clinical and environmental data for use in clinical research although its utility for clinical and genomic research is still being explored.

Similarly, the mapping of the human genome has enabled new exploration of how genetic variations contribute to health and disease. Genomic research has been very successful with the results of this work both shedding light on how genetic variants influence susceptibility to common, chronic diseases but also playing an instrumental role in the discovery of new biologic pathways and drug targets. The combination of genotypic data and phenotypic data from EHRs is a pivotal first step towards the implementation of personalized medicine in clinical care. Advances in genetics research are driving a need for ever larger sample sizes and a possible solution has been the establishment of biorepositories linked to EHRs, and EHRs are playing an important role in this. Several national initiatives have formed to explore the strengths and limitations of EHRs as a source of phenotypic data. The Electronic Medical Records & Genomics Network (eMERGE) is a consortium formed in 2007 by the National Human Genome Research Institute (NHGRI) and now contains 10 institutions.[1] The Strategic Health IT Advanced Research Project for data normalization (SHARPn) was a collaboration about 17 academic institutions and industry that developed strategies and methodologies for large-scale EHR data standardization and scalable phenotyping to improve the overall quality of healthcare.[2] The Pharmacogenomics Research Network (PGRN) and the PGRN resource, Pharmacogenomics in Very Large Populations (PGPop), have also used EHR data for genomic research. Like eMERGE, PGPop has executed drug-response phenotypes across multiple sites. eMERGE has studied >40 phenotypes with genome-wide associations studies, noting a number of new discoveries. More recently, the Patient Centered Outcomes Research Institute (PCORI) created PCORnet, a national research network of 11 Clinical Data Research Networks and 18 Patient Powered Research Networks to conduct comparative effectiveness research on topics of direct importance to patients. The PCORnet distributed research model will necessitate computable phenotypes that can be executed across its component networks, and also anticipates linkage of EHR data to biorepository data to enable genomic research.[3] PCORnet's approaches to phenotyping and research - healthcare

system collaboration have been heavily influenced by the NIH Health Care Systems Research Collaboratory, a demonstration program for the transformation of clinical trials based upon use of electronic health records (EHRs), healthcare systems partnerships.[4] The Collaboratory has developed guidelines for the identification, evaluation, and use of EHR-based phenotypes in pragmatic clinical trials.

Phenotype algorithms in eMERGE and other networks have been shared via PheKB.org. To harmonize data across sites, eMERGE institutions and collectively have also pioneered the use of phenome-wide associations studies (PheWAS); these results have also been shared. The goal of the network is to develop, disseminate and apply research methods that combine EHRs with DNA biorepositories for the conduct of large-scale genome-wide association studies. These institutions have developed and validated a variety of phenotypes using the EHR. eMERGE has since launched a prospective next-generation sequencing project, eMERGE-PGx, which is using phenotype algorithms to select patients most likely to be exposed to medications whose efficacy is influenced by pharmacogenomics. This effort, coupled with computerized decision support, will sequencing ~9000.

Topic

The panel represents data and informatics experts from the eMERGE, PGRN, PGPop, SHARPn, NIH Collaboratory, and PCORnet consortia who will share their experiences in the use of EHR data for genetic and clinical research. Specifically each panelist will present the performances and challenges of implementing locally and externally developed electronic phenotype algorithms, efforts to create computational algorithms, and novel approaches to computational phenotyping algorithms, including PheWAS approaches, topic modeling, and non-negative tensor factorization. We will also discuss PheKB.org, phenotype standardization, and efforts to harmonize terminologies and data dictionaries. The panel will also include structured discussions on the following topics:

Phenotyping Applications and Experiences

- Goals and aspirations of each of the involved networks, and experiences in each
- Genomic and pharmacogenomic studies results from each site, and the incorporation of new sites
- Discussion of experiences with cross-institution implementation of phenotype algorithms and local adaptations required for the network-wide phenotypes
- Evaluation of the accuracy of different EHR categories of information for accurate phenotype algorithms
- Phenome-wide association studies (PheWAS) using EHR data for relevant genomic variants

Phenotyping Methodologies

- Development of sharable phenotype libraries available for public use (PheKB.org)
- Models for using phenotype libraries by different research networks and consortia
- Novel computing methods for high-throughput phenotyping
- Methodologies for phenotyping refinement and validation
- Methodologies for adapting phenotyping algorithms across sites
- Representation of phenotype data using emerging data standards

Panel Participants

Joshua Denny – Vanderbilt University – Moderator
Dr. Denny is an Associate Professor in the Department of Biomedical Informatics at Vanderbilt University, and was the SPC chair of the 2014 Summit on TBI. Vanderbilt uses an opt-out model biobank associated with a de-identified Synthetic Derivative[5] of the EHR. Vanderbilt's has used BioVU for more than 100 studies, including investigation of >30 pharmacogenetic phenotypes led by Dr. Denny. Individuals were selected amongst those without cardiac disease, and analysis with and without evidence of possibly-interfering medications. Our results identified genomic variants associated with cardiac conduction (e.g., the QRS and PR intervals[6] on electrocardiogram), validating use of EHR-based biobanks for genomic studies. He also developed algorithms and software to perform phenome-wide association analyses (PheWAS), scanning the EHR phenome for genetic associations with many diseases not anticipated in the initial GWAS, and use of this tool to explore pleiotropy in current GWA studies. Dr. Denny has also led development of PheKB.org to share phenotype algorithms and phewascatalog.org to share PheWAS results.

Rachel Richesson – Duke University
Dr. Richesson is an Associate Professor of Informatics at the Duke University of Nursing, and was SPC chair of the 2014 Summit on CRI. She is particularly interested in applications and standards specifications that increase the efficiency of clinical research and enable interoperability between clinical research and health care systems. She co-

leads the Phenotyping, Data Standards, and Data Quality Core for the NIH Health Care Systems Research Collaboratory, a demonstration program for the transformation of clinical trials based upon use of electronic health records (EHRs) and healthcare systems partnerships. In this role, she is developing standard approaches and guidance for using computable phenotypes in the extraction of clinical data to support research and learning healthcare. She is also the co-lead of the Rare Diseases Task Force for the national distributed Patient Centered Outcomes Research Network (PCORnet), specifically promoting standardized computable phenotype definitions for rare diseases, and helping to develop a national research infrastructure that can support observational and interventional research for various types of conditions.

Abel Kho – Northwestern University
Dr. Kho is an Assistant Professor of Medicine in General Internal Medicine at Northwestern University and a member of the phenotype working group of eMERGE and Co-Chair of the PCORnet Data Standards, Security and Network Infrastructure Task Force. Northwestern's GWAS assessed genomic variants associated with Type 2 Diabetes (T2D) and more recently diverticulosis and diverticulitis. Dr. Kho will compare phenotyping experiences on eMERGE and PCORnet.

Jyotishman Pathak – Mayo Clinic
Dr. Pathak is an Associate Professor of Biomedical Informatics at the Mayo Clinic College of Medicine. His research focuses on developing methodology and applications for standardized and computable phenotyping algorithms. Dr. Pathak will describe recent progress from the SHARPn collaboration developing open-source solutions for high-throughput and scalable phenotyping using Hadoop-based big data technologies, and highlight on-going work in creating a robust platform for phenotyping algorithm authoring and dissemination. Dr. Pathak has also led the development of phenotype data harmonization platform within eMERGE and PGRN.

Jimeng Sun – Georgia Institute of Technology
Jimeng Sun is an Associate Professor of School of Computational Science and Engineering at College of Computing in Georgia Institute of Technology. Prior to joining Georgia Tech, he was a research staff member at IBM TJ Watson Research Center. His research focuses on health analytics using electronic health records and data mining, especially in designing novel tensor analysis and similarity learning methods and developing large-scale predictive modeling systems. Dr. Sun will describe a recent collaborative project on computational phenotyping using tensor factorization[7,8]. The goal of the project is to develop algorithms and systems for automatically converting EHR data into meaningful clinical concepts (phenotypes) that can be used for clinical research and clinical decision making. The proposed algorithms and methods in this project are expected to enable 1) phenotype generation and refinement within an institution, 2) adaptation across institutions and 3) applications in a broad range of health informatics such as GWAS and predictive modeling.

All participants have agreed to take part on the panel.

References

1. Gottesman O, Kuivaniemi H, Tromp G, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet. Med.* 2013. doi:10.1038/gim.2013.72.
2. Chute CG, Pathak J, Savova GK, et al. The SHARPn Project on Secondary Use of Electronic Medical Record Data: Progress, Plans, and Possibilities. *AMIA Annu Symp Proc* 2011;2011:248-256.
3. Collections, PCORNet. Available at: http://jamia.bmj.com/collections/pcornet. Accessed September 25, 2014.
4. Richesson RL, Hammond WE, Nahm M, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* 2013;20(e2):e226-e231. doi:10.1136/amiajnl-2013-001926.
5. Roden DM, Pulley JM, Basford MA, et al. Development of a large-scale de-identified DNA biobank to enable personalized medicine. *Clin. Pharmacol. Ther* 2008;84(3):362-369. doi:10.1038/clpt.2008.89.
6. Denny JC, Ritchie MD, Crawford DC, et al. Identification of genomic predictors of atrioventricular conduction: using electronic medical records as a tool for genome science. *Circulation* 2010;122(20):2016-2021. doi:10.1161/CIRCULATIONAHA.110.948828.
7. Ho, Joyce C., Joydeep Ghosh, Steve R. Steinhubl, Walter F. Stewart, Joshua C. Denny, Bradley A. Malin, and Jimeng Sun. "Limestone: High-Throughput Candidate Phenotype Generation via Tensor Factorization." *Journal of Biomedical Informatics*. Accessed August 8, 2014. doi:10.1016/j.jbi.2014.07.001.
8. Ho, Joyce C., Joydeep Ghosh, and Jimeng Sun. "Marble: High-Throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization." In KDD '14.

Panel: Challenges of Implementing Genomic Decision Support in the Real World—Experience from the eMERGE and CSER Networks.

Justin Starren[1], Brian Shirts[2], Peter Tarczy-Hornoch[2], Marc Williams[3], Tim Herr[1]
[1]Northwestern University Feinberg School of Medicine, Chicago, IL; [2]University of Washington School of Medicine, Seattle, WA; [3]Geisinger Health System, Danville, PA

Abstract

While the need for genomic decision support has been universally recognized and discussed in numerous papers, the number of actual implementations has been few. Many commercial EHR vendors have stated publicly that they do not have near-term plans to store genomic data within the EHR itself, although they anticipate exposing genomic computerized decision support (GCDS) within their workflows. Several papers have explicitly commented on the need for Ancillary Genomic Systems (AGS) to mediate between the large volumes of genomic data that is anticipated and the limited data capacity of most EHR systems. The Electronic Medical Records and Genomics (eMERGE) network and the Clinical Sequencing Exploratory Research (CSER) network are to multi-institution networks exploring the challenges of implementing GCDS in real-world situations. This panel brings together representatives of both networks to discuss the collective experience of 16 different health systems that have incorporated genomic results into the process of care.

Panel Overview

The eMERGE consortium has previously presented at AMIA about its plans for integration of CGDS into the EHR workflow. All 10 eMERGE sites are now live with GCDS, and have recently completed an analysis of the challenges encounter is implementing these systems. Each site in the CSER consortium has implemented methods for return of sequencing results to patients. Recently, the two consortia have pooled their expertise and experience with integration of genomic results into EHRs. This panel presents that combined experience. To rapidly address these questions, the National Human Genome Research Institute (NHGRI) and the National Cancer Institute (NCI) have initiated a Clinical Sequencing Exploratory Research (CSER) program to support multidimensional research in this area. CSER is a national consortium of projects that bring together clinicians, scientists, laboratories, bioinformaticians, economists, legal scholars, ethicists, and patients working together to develop and share innovations and best practices in the integration of genomic sequencing into clinical care.

Each consortium has studied the experience of its member sites in incorporating genomic results into clinical care. These studies have identified a wide variety of challenges, some that have been discussed previously, and many that are either novel or underappreciated. In addition, the two networks are collaborating on a joint analysis of the incorporation and presentation of genomic information in various EHR systems. In addition, the two consortia are jointly developing recommendation for optimal presentation of genomic information in conventional EHRs.

By bringing together these two consortia, this panel provides an unusually broad view of the domains. Although members of these consortia have presented independently at AMIA meetings regarding ongoing activities, the consortia members have now gone live with their implementations. As a result, this panel represents the first summative presentation of implementation experience across the two consortia.

Justin Starren — Moderator

Dr. Starren is co-Chair of the EHR Integration Workgroup of eMERGE. In addition to his role in eMERGE, Dr. Starren is the informatics liaison between the eMERGE and CSER consortia. This bridging position makes him ideally suited to bring together the various informatics activities across the two consortia.

Peter Tarczy-Hornoch—Integration of Clinical Sequencing data into the EHR.

Dr. Tarczy-Hornoch is Co-PI of the Clinical Sequencing Exploratory Research (CSER) Coordinating Center focused on CSER informatics issues, past Chair of the CSER EHR working group, chaired the 9th AMIA Policy Invitational "Harnessing Next-Generation Informatics for Personalizing Medicine" in 2014 and has a long standing interest in genomic data integration, annotation and curation. He has previously published on the variability of methods for integration of clinical sequencing data, even in the presence of similar underlying EHR infrastructures across CSER

sites. He will present an overview of the CSER sites with an emphasis on the informatics aspects illustrated with examples from prior CSER EHR working group activities and site-specific activities.

Brian Shirts – Presentation of genomic data within the EHR

Genomic information comes in many forms and has many different diagnostic and therapeutic uses. Consequently, it may be presented in many different part of an EHR system. Moreover, individual sites may differ on where within the EHR that particular types of genomic information are stored and presented. Dr. Shirts is a Molecular Genetics Pathologist and clinical laboratory informaticist. He is leading a joint CSER/eMERGE project exploring where and how different categories of genomic information are currently presented in EHRs aimed at determining if the presentation of genetic information in the EHR is facilitating the many ways that clinicians can use this information to benefit patients. The project is also developing recommendations on how EHR systems can improve to allow optimal display of genetic information and accompanying decision support.

Marc Williams -- Vanderbilt

Marc Williams is a clinical geneticist and co-chair of the eMERGE EHR Integration workgroup. He will provide an overview of the current state of the eMERGE GCDS implementation as well as placing the eMERGE activities within the national context. In addition to his work in eMERGE, he is involved in multiple other national initiatives related to integration of genomic data in clinical care. He participated in national efforts in this regard including the American Health Information Community's Personalized Healthcare Workgroup, which published the recommendations for a minimum data set of family history for electronic health records and addressed gaps in the representation and communication of genomic data in electronic health records. Individuals from my team at Intermountain Healthcare participated in the creation and adoption of the Health Level Seven (HL-7) genomic standards. While at Intermountain he led a team that worked in conjunction with members of the Homer Warner Center for Informatics Research and the Interface team that performed the first successful transmission of a molecular genetic test result from a referral laboratory into the electronic health record. He was a member of the planning team and co-chaired the NHGRI-sponsored meeting that developed the ClinGen RFA and am lead author on a paper describing the important issues that were identified at the meeting. He is investigator on the multi-site ClinGen project with the specific task of studying the integration of the ClinGen resource with certified EHRs. He is a member of the Epic Genomic Medicine workgroup that is exploring how to incorporate genetic and genomic information into the Epic electronic health record. He also led the NHGRI's participation in the NIH Dissemination and Implementation effort culminating in a workshop on implementation of individualized medicine at the 5th annual meeting on Dissemination and Implementation in March of 2012. He is also a member of the Clinical Decision Support Consortium 2.0 and the Clinical Pharmacogenetics Implementation Consortium informatics workgroup.

Tim Herr—Survey of Impact of Genomic CDS Implementation

Tim Herr is a PhD student in Health and Biomedical Informatics at Northwestern University. He is leading an effort to collect data across the eMERGE network to analyze specific implementation challenges related to the genomic decision support projects at each of the ten eMERGE sites. Tim will present on the lessons learned from this project. The paper based on this analysis is in final preparation and we expect this work to be in press by the time of the Joint Summits. Previously, Tim built expertise in the management and analysis of electronic health records and public health databases in industry positions at Epic and Hewlett-Packard.

Affirmation

This is to affirm that all proposed members have been personally contacted by Dr. Starren and have agreed to participate in this panel.

Adverse Drug Event Ontology: Gap Analysis for Clinical Surveillance Application

Terrence J. Adam, RPh, MD, PhD[1,2], Jin Wang, MSPharm, MHI,[1,2]
[1]Institute for Health Informatics; [2]College of Pharmacy, University of Minnesota, Minneapolis, MN.

Abstract: *Adverse drug event identification and management are an important patient safety problem given the potential for event prevention. Previous efforts to provide structured data methods for population level identification of adverse drug events have been established, but important gaps in coverage remain. ADE identification gaps contribute to suboptimal and inefficient event identification. To address the ADE identification problem, a gap assessment was completed with the creation of a proposed comprehensive ontology using a Minimal Clinical Data Set framework incorporating existing identification approaches, clinical literature and a large set of inpatient clinical data. The new ontology was developed and tested using the National Inpatient Sample database with the validation results demonstrating expanded ADE identification capacity. In addition, the newly proposed ontology elements are noted to have significant inpatient mortality, above median inpatient costs and a longer length of stay when compared to existing ADE ontology elements and patients without ADE exposure.*

Introduction: Adverse drug events, including injury and death, have been reported to affect up to 1.6 million patients annually, according to an Institute of Medicine report[1] posing an important clinical problem. Adverse drug events are a major patient safety concern with the Centers for Medicare and Medicaid Services developing early efforts to consider mandating adverse event monitoring and subsequent development of the Sentinel Initiative to facilitate medication safety efforts[2]. Exact numbers are difficult to assess, due to lack of uniform reporting, but adverse drug events add 2-4 days to hospital length of stay at a cost of $2500-$5500 per patient[3-5].

Although adverse drug events may occur frequently, the ability to identify actual adverse events has been difficult due to the limited availability of effective methods. Voluntary reporting of inpatient adverse drug events are the most well-known systems for event tracking, but are the poorest performing of the systems, identifying only around 1 in 20 events due to insufficient provider participation[6]. Given the difficulty in identifying clinical events due to the limited availability of well curated and standardized electronic medical record data and the lack of effective reporting by providers, automated surveillance methods are needed for population medication safety surveillance. Given that automated system detection rates can be three to twelve times higher than provider driven reporting there is substantial potential to fill this critical safety surveillance gap[7].

Background: The inpatient clinical population has a higher rate of adverse drug events given the higher level of clinical acuity and underlying clinical comorbidities requiring hospitalization[8]. Adverse drug events are difficult to detect using chart reviews due to resource cost and data quality limitations resulting in challenges with data aggregation for population risk assessment of low prevalence adverse events. For clinical sites with advanced electronic medical record capacity and clinical data review resources, there may be enhanced capabilities to identify local clinical events; however, there may not be a sufficient number of drug exposures to effectively identify adverse drug event patterns. In order to maximize the likelihood of detecting adverse drug events, large clinical databases with capacity for efficient data aggregation can more effectively detect rare adverse drug events and accurately assess the risk of adverse events in populations. Efficient ADE identification can also facilitate medication safety initiatives, medication outcomes assessment and drug-drug interaction risk assessment. Pairing a comprehensive ADE ontology with a large clinical database such as the Healthcare Cost and Utilization Project (HCUP) National Inpatient Sample (NIS) can provide important insights on events rates, ADE trends and event predictors at the facility level given that the data includes a 20% US hospital sample stratified for region, location, teaching status, bed size and ownership http://www.hcup-us.ahrq.gov/nisoverview.jsp. Given the critical need for effective, low- cost and computational driven approaches to ADE surveillance, an up-to-date and comprehensive ADE ontology can help identify ADEs present in structured clinical data both to better understand ADE incidence, facilitate quality improvement initiatives, and manage ADE surveillance.

Methods

Ontology Development. A comprehensive ADE ontology was developed from available clinical literature, ADE public use files and a large clinical database using a Minimum Clinical Data Set (MCDS) Framework[9]. The available literature and ADE public use files were assessed for both ADE identification capabilities and ADE surveillance gaps with each source being iteratively incorporated into the final comprehensive candidate ontology.

Relevant ICD-9 CM diagnostic and E-codes in the current ICD9- CM clinical code set were identified using text searches for candidate identification followed by confirmation with expert review. Two comprehensive public use ADE code lists were analyzed including the 2011 Healthcare Cost and Utilization Project (HCUP-ADE) (http://www.hcup- us.ahrq.gov/reports/statbriefs/sb158_ADE_Appendix.pdf) and 2002 Utah/Missouri Patient Safety Project (UMPSP) (http://health.utah.gov/psi/icd9.htm) with supplementation from prior studies on structured data for ADE identification [8, 10, 11].

Ontology Validation. Using the MCDS approach, the proposed ontology was applied to relevant clinical data to assess viability and impact. Each proposed code was assessed for its application in the clinical data (code viability) as well as its association with clinical outcome (clinical impact). Candidate codes which were not present in the clinical dataset were removed from the candidate list. The proposed new codes were grouped into two parent categories: Administrative ADE and Other ADE medications. The mortality rates, average length of stay and average total charges were assessed for both parent categories with the results summarized in Table 2.

Results: The HCUP-ADE code set containing 467 ADE codes in 36 therapeutic categories provided the best currently maintained public use file for initiating the comprehensive ontology development and was the gold standard for the study evaluation. The final comprehensive ADE ontology contained a total of 531 ADE codes after completing the iterative incorporation of source content. The final ontology also added two additional categories including "Administrative ADE" related to adverse events related to medication administration with the remaining additional codes being grouped in the "Other ADE medications" for validation purposes. The "Administrative ADE" group contained a total of 17 proposed ADE codes and the "Other ADE Medications" included a total of 47 proposed ADE codes for a total of 64 new codes proposed for inclusion in conjunction with the current HCUP-ADE framework containing 467 clinical codes.

Ontology Validation: O**verall ADE Rates:** The total number of hospitalization discharges reviewed from the 2011 HCUP National Inpatient Sample (NIS) database included 8.024 million hospitalization events from a stratified probability sample of hospitals, with sampling probabilities calculated to select 20% of the universe of U.S. community, non-rehabilitation hospitals contained in each sample stratum. Using the HCUP-ADE public use file alone for gold standard benchmarking, a total of 735,050 ADE events were identified. After incorporating the two new proposed ADE code sets into the expanded comprehensive ontology, a total of 905,001 events were identified providing a 23.1% increase in ADE identification. Since each hospitalization could have more than one ADE, the number of hospitalizations with one or more ADE were identified yielding a total of 639,884 unique ADE associated hospitalizations producing a 7.98% ADE hospitalization rate.

Subject Characteristics and Overall Outcomes: In table one, the characteristics of the patients with and without ADEs are summarized and the groups are significantly different for each measure including a 66.7% relative increase in inpatient mortality, a 2.4 day increase in length of stay and $18,827 increase in average cost among patients experiencing an ADE. The ADE>0 group was older with more chronic diseases and had a higher percentage of Medicare patients and a lower numbers of Medicaid and private pay patients ($P < 0.01$). The ADE group also had slightly higher income and higher proportion of Caucasians ($P < 0.01$).

Table 1: ADE Patients and Non-ADE Patients

	ADE=0	ADE>0	p-value
Age (sd)	48.7±28	59.4±21	p<0.0001
Length of Stay(sd)	4.4±6.5	6.8±9.4	p<0.0001
Total Charges (sd)	33963±61404	52790±99245	p<0.0001
# of Chronic Diseases	4±3.5	6±3.3	p<0.0001
Female (%)	58.2	56.4	p<0.0001
Mortality (%)	1.8	3.0	p<0.0001

ADE Therapeutic Group Rates and Association with Outcomes: The medication group event rates, mortality, length of stay (LOS) and total charges are noted in table 2 to provide an evaluation of the association of the type of ADE with outcomes. The results from the proposed expanded ontology are included in the "Administrative ADE" (new) and the "Other ADE medications" (new) categories with the remainder of the groups from the HCUP gold standard.

Table 2: ADE Therapeutic Group Event Rates and Outcomes:							
ADE Group	**Events**	**Mortality**		**LOS**		**Total Charges**	
		Freq	**Rate**	**Mean**	**Std Dev**	**Mean**	**Std Dev**
Non-ADE patients	7383706	132872	1.8	4.4	6.5	33963	61404
Antibiotics	30075	620	2.07	8.5	10.7	63728	116910
Clostridium difficile infection	79633	6195	7.79	11.6	14.8	88423	166622
Other anti-infectives	9650	164	1.7	6.5	10.6	44597	97223
Steroids	64804	1678	2.6	7.1	8.2	58664	98423
Insulin and Hypoglycemics	9991	133	1.34	4.4	7.7	30911	49320
Other hormones	3984	33	0.83	4.3	6.3	31956	52339
Antineoplastic drugs	62025	2957	4.78	7.8	9.3	62186	103685
Anti-allergy and antiemetic drugs	4784	35	0.73	4.3	6.0	31063	59302
Other systemic agents	1122	20	1.79	4.7	7.5	35579	59829
Anticoagulants	49908	2582	5.19	6.8	7.2	55395	84846
Other agents affecting blood constituents	6755	299	4.43	6.6	7.9	65607	102534
Opiates/Narcotics	36149	698	1.94	5.5	6.9	46970	72129
NSAIDS	40643	655	1.62	4.5	5.7	38029	61957
Hydantoin	4407	65	1.48	6.0	8.3	41433	91714
Other anticonvulsants	10083	78	0.78	5.3	8.3	34470	67551
Anti-Parkinson drugs	1078	8	0.74	5.5	6.5	29971	37292
Sedatives or hypnotics	21044	347	1.65	5.8	7.3	50263	78422
CNS depressants and anesthetics	10832	139	1.29	4.5	7.0	44592	84622
Antidepressants	12990	91	0.7	4.0	6.2	25504	49983
Antipsychotics	9745	90	0.93	6.3	11.1	33473	67547
Benzodiazepine	25118	341	1.36	4.1	6.2	31150	59103
Other psychotropic drugs	9812	125	1.28	5.5	9.1	33206	53089
Central nervous system drugs	6706	133	1.99	4.0	6.5	34601	61335
Autonomic nervous system drugs	3876	74	1.91	5.4	7.3	39578	72908
Digoxin	6508	344	5.29	6.2	6.9	45973	71764
Anti-adrenergics	6695	62	0.93	4.1	5.2	33284	47013
Other cardiovascular drugs	32323	408	1.27	4.8	6.4	38444	68210
GI system drugs	2268	36	1.59	5.7	6.4	42827	64556
Saluretics	7464	57	0.77	4.2	5.0	32404	44814
Other diuretics	17810	429	2.42	6.0	7.8	46836	88445
Other mineral and uric acid metabolism	3106	96	3.1	7.5	11.6	60064	141985
Smooth muscle and respiratory drugs	6084	34	0.56	4.1	5.8	28574	43714
Skin, eye, mucous membrane drugs	1561	16	1.03	5.9	8.7	40979	70417
Vaccines	465	4	0.86	4.2	4.9	29106	37750
Other specific drugs	237	3	1.27	3.3	4.4	27092	44596
Nonspecific ADE causes	79556	2133	2.68	7.2	9.8	56057	98530
Administrative ADE (new)	**2268**	**114**	**5.03**	**8.0**	**10.0**	**102852**	**170667**
Other ADE medications (new)	**148931**	**2055**	**1.38**	**5.8**	**8.5**	**43266**	**84427**

The proposed "Administrative ADE" and "Other ADE medications" groups appear to have clinical and economic importance. The proposed "Administrative ADE" group had the fourth highest mortality, third longest length of stay (LOS), and the highest average hospital cost. The proposed "Other ADE medications" had the 22nd highest mortality, the 16th longest LOS and 16th highest average cost among the 39 therapeutic areas providing data on the relative ADE impact and evidence for ontology inclusion.

Code Viability: Each of the 531 codes were assessed for their presence in the 2011 NIS database to provide evidence of content use and viability with the assumption that missing codes in a large clinical database may be antiquated medications or the ADE code may have been changed to a different code. Among the 64 proposed new ADE codes, 62 codes were identified in the database with at least one clinical instance providing a coverage rate of 96.9%. The new codes which were not identified included the "Failure of sterile precautions during infusion or transfusion" code which was in the "Administrative ADE medications" group and "Psychostimulant poisoning" which was in the "Other ADE medications" category. The gold standard HCUP data set was also evaluated with 425 of the 467 ADE codes being noted in the NIS database yielding a coverage rate of 91.0%. The bulk of the codes which were not found in NIS were in the vaccine group (15 codes), CNS depressants and anesthetics (9 codes) and sedative/hypnotic (8 codes), with the other potentially non-viable ADE codes in other categories. Combining the 467 gold standard HCUP-ADE codes with our proposed 64 code addition yielded a composite coverage of 91.7%.

Discussion: The proposed comprehensive ADE ontology provides higher levels of ADE identification than the HCUP-ADE dataset with a greater than 22% expansion in identified events. Given that the proposed ontology expansion groups had significant inpatient mortality, average cost and length of stay, the inclusion of these coded elements appear to be clinically and economically meaningful. The NIS data validation step provided evidence for the need for an expanded structured ADE ontology given the substantially higher ADE rate. Our identified event rates are substantially higher than comparable studies with HCUP-ADE which yielded an ADE rate of 5.3% for Medicare patients[8] and 5.64% for Stausberg[10] on an earlier but similarly broad NIS data set from 2006. Though our findings had a higher ADE rate than previous publications on older NIS data,[8, 10, 11] this may be partially related to the secular trend of increased ADE events noted by Shamliyan[8] where ADE rates have increased by 90% from 2000 to 2008 in Medicare specific patients. Further expansion of the ADE ontology may or may not affect the event trend, but the results warrant addition assessment with longitudinal data to better understand the event rate and trend.

The events in the new "Administrative ADE" group had particularly high clinical impact and substantial economic cost providing important evidence for ADE inclusion. The "Other ADE medications" group did not have as high of inpatient mortality as the "Administrative ADE" group but was near the median in terms of inpatient mortality among the established gold standard HCUP-ADE therapeutic groups and was above the median HCUP-ADE therapeutic group LOS and average cost supporting their inclusion in a comprehensive ontology.

The "Other ADE medications" list included a diverse set of ADE codes with a number of therapeutic areas which may not be best represented as a separate category. Two therapeutic areas which were in the new "Other ADE medications" group were psychiatric and pain medications. They may be incorporated into the existing HCUP-ADE Opiates/Narcotics, Antipsychotics, Antidepressants or Other psychotropic drug groups to better cluster the affected patients and events since their inclusion in the "Other ADE medications" list likely overlaps, at least partially, with those existing HCUP-ADE categories. Several of the other ADE codes in the "Other ADE medications" group did not appear to map to existing HCUP-ADE therapeutic groups such as those for drug dermatitis and drug events among obstetrics and newborns. Further exploration of potentially new HCUP-ADE therapeutic categories may be warranted for the obstetrics and newborn category while many of the other ADE may be managed on an interim basis in the existing HCUP "Other specific drugs" and "Nonspecific ADE causes" ADE categories. The proposed "Administrative ADE" category would likely be a viable new addition to the HCUP-ADE since it is clinically and economically important and reflects clinical events which are by and large preventable using appropriate safety measures and are without a close fitting category in the current HCUP-ADE data set.

With the existing HCUP-ADE data set, a substantial number of ADE codes were not found in NIS validation including codes related to older and rarely used medications such as arsenic anti-infectives and mercurial anti-diuretics which have been replaced in clinical use with less toxic alternative medications. Also missing from NIS were a number of anesthesia related events and adverse events associated with vaccinations which is in part due to low levels of immunizations for diseases such as smallpox and yellow fever which are not on the usual care pathway for childhood and adult vaccinations resulting in relatively low numbers of exposed patients and associated ADE. Further exploration of other years of NIS data and other clinical data sources may be needed to assess if thesemissing codes are used in older data, other clinical settings or geographic locations where exposure rates are higher due to differences in disease prevalence and treatment care standards.

The study results have a number of important limitations. The NIS data does not distinguish drug associated harms which occur in the outpatient setting leading to hospitalization from those which occur in the inpatient setting. In addition, the NIS data has a number of other deficiencies, most notably, the lack of a full medication profile which limits the ability to associate ADE to exposure medications. The NIS data sampling is based on a hospital sample which may not accurately represent all inpatient admissions, however, in 2012; the NIS dataset methodology was changed to better reflect event rates at the patient admission level rather than facility level which may better address the true event rates.

In future work, the incorporation of specific medication data at the individual level using electronic medical record, Medicare, Medicaid or commercial claims data would provide additional insight on ADE risk factors and their association with clinical outcomes. Such data would be important to increase the granular understanding of the problem as well as to explore ADE risk prevention strategies. The comprehensive ADE ontology may also be used for retrospective ADE identification for targeted risk assessment and mitigation since it uses administrative data which is operational in most US clinical settings with the capacity to quickly identify problem areas and to benchmark against national average data noted in Table 2. This rapid potential to translate into retrospective surveillance could provide an important tool to use alongside data mining, natural language processing and epidemiologic methods to operationalize drug safety initiatives and research.

References:

1. Preventing Medication Errors: Quality Chasm Series. Aspden P, Wolcott J, Bootman JL, Cronenwett LR, editors: The National Academies Press; 2007.
2. FDA. FDA's Sentinel Initiative2011 04/22/11. Available from: http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm.
3. Bates DW, Spell N, Cullen DJ, Burdick E, Laird N, Petersen LA, et al. The costs of adverse drug events in hospitalized patients. Adverse Drug Events Prevention Study Group. JAMA : the journal of the American Medical Association. 1997;277(4):307-11. Epub 1997/01/22.
4. Suh DC, Woodall BS, Shin SK, Hermes-De Santis ER. Clinical and economic impact of adverse drug reactions in hospitalized patients. The Annals of pharmacotherapy. 2000;34(12):1373-9. Epub 2001/01/06.
5. Classen DC, Pestotnik SL, Evans RS, Lloyd JF, Burke JP. Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA : the journal of the American Medical Association. 1997;277(4):301-6. Epub 1997/01/22.
6. Jha AK, Kuperman GJ, Teich JM, Leape L, Shea B, Rittenberg E, et al. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. J Am Med Inform Assoc. 1998;5(3):305-14. Epub 1998/06/03.
7. Kilbridge PM, Campbell UC, Cozart HB, Mojarrad MG. Automated surveillance for adverse drug events at a community hospital and an academic medical center. J Am Med Inform Assoc. 2006;13(4):372-7. Epub 2006/04/20.
8. Shamliyan TA, Kane RL. Drug-Related Harms in Hospitalized Medicare Beneficiaries: Results From the Healthcare Cost and Utilization Project, 2000-2008. Journal of patient safety. 2014. Epub 2014/06/01.
9. Svensson-Ranallo PA, Adam TJ, Sainfort F. A framework and standardized methodology for developing minimum clinical datasets. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science. 2011;2011:54-8. Epub 2012/01/03.
10. Stausberg J. International prevalence of adverse drug events in hospitals: an analysis of routine data from England, Germany, and the USA. BMC health services research. 2014;14:125. Epub 2014/03/14.
11. Shamliyan T. Adverse drug effects in hospitalized elderly: data from the healthcare cost and utilization project. Clinical pharmacology : advances and applications. 2010;2:41-63. Epub 2010/01/01.

Integrating Gene Regulatory Networks to identify cancer-specific genes

Valeria Bo[1] and Allan Tucker[1]

[1]Department of Computer Science, Brunel University, London, UK

Abstract

Consensus approaches have been widely used to identify Gene Regulatory Networks (GRNs) that are common to multiple studies. However, in this research we develop an application that semi-automatically identifies key mechanisms that are ***specific*** *to a particular set of conditions. We analyse four different types of cancer to identify gene pathways* ***unique*** *to each of them. To support the results reliability we calculate the prediction accuracy of each gene for the specified conditions and compare to predictions on other conditions. The most predictive are validated using the GeneCards encyclopaedia[1] coupled with a statistical test for validating clusters. Finally, we implement an interface that allows the user to identify unique subnetworks of any selected combination of studies using AND & NOT logic operators. Results show that unique genes and sub-networks can be reliably identified and that they reflect key mechanisms that are fundamental to the cancer types under study.*

Introduction

When an organism is subjected to a different condition either internal or external to it (environmental changes, stress, cancer, etc.) its underlying mechanisms undergo some changes. To build robust and reliable Gene Regulatory Networks (GRNs) from microarrays, it is necessary to integrate multiple data collected from different studies[2,3,4,5]. To identify links in common among a set of independent studies, researchers apply consensus networks analysis. Swift et al.[6] apply a clustering technique coupled with a statistically based gene functional analysis for the identification of novel genes. While Segal et al.[7] group genes that perform similar functions into *'modules'* and then build networks of these modules to identify mechanisms at a more general (higher) level. More recently, a similar approach[8] was applied to a large number of cancer datasets where case and control are compared. For each dataset, the pairwise correlation of gene expression profile is computed and a frequency table is built. Then the values in the table are used to build a weighted gene co-expression frequency network. After this they identify sub-networks with similar members and iteratively merge them together to generate the final network for both cancer and healthy tissue.
In[9], we expand on this work but rather than focusing on consensus networks, we develop a method to 'home in' on both the similarities and *differences* of GRNs generated from different independent studies by using a combination of partial correlation network building and graph theory. The method goes beyond the simple pairwise correlations between genes, as in[8], by building independent networks for each study using *glasso* which identifies the inverse covariance matrix using the lasso penalty. Rather than identifying consensus studies, we detect the edges that are $unique/specific$ for each study and build Bayesian Networks to identify the most predictive group of genes and further refine our networks.
In this work we extend the work presented in[9] by exploring the performances of the pipeline using four different cancer datasets and identifying, through the GeneCards encyclopaedia[1], the list of genes known to be involved in each type of cancer. We apply a statistical test to measure the significance of detecting these genes in our unique networks. In addition, we develop an interface that allows the user to select combinations of studies using AND and NOT logic operators and to identify the related unique sub-networks and genes.

Materials and Methods

In this paper we adapt the Unique Network Identification Pipeline (UNIP) developed in[9]. Each step of the pipeline applied for the specific case of this paper are explained in the following sections.

Dataset Description. Four different cancer datasets are downloaded from the NCBI Gene Expression Omnibus (GEO) website[10]. To avoid platform bias the datasets selected are all generated using Affymetrix HU133 Plus 2.0 Genechip. Given the raw series of data, the *rma* (Robust Multi-Array Average) expression measure (available in the R package 'affy'[11]) is applied as a pre-processing step. Each study identification code, description and samples number are summarized in Table 1.

i) Selection of Informative Genes. The high discrepancy between the number of genes (order of thousands) and the samples (tens or hundreds) measured simultaneously in microarray data leads to the necessity of reducing the number of variables (genes) involved in the analysis. R statistics provides the 'pvac' package[12] which applies the PCA (Principal Component Analysis)[13] and returns a subset of the original variables: the closest to the principal components identified. To further refine the variable reduction and to select the most active genes, the standard deviation of each gene across all the samples in each separate study is calculated and only genes with $sd \geq 1.5$ in at least one of the 4 studies are selected. The reduced datasets are used as input to the following steps of the analysis.

ii) Glasso. At this stage we need to build a GRN for each condition/study in the dataset.
As we want to identify networks that go beyond simple pairwise relationships, our procedure uses *glasso*[14,15,16], which calculates the inverse covariance matrix using the lasso penalty to make it as sparse as possible.
In this paper, we apply *glasso* with the penalization parameter $\rho = 0.05$, to build a GRN for each study dataset. In addition, to further improve the sparsity and reduce the nodes involved, we maintain only the connections with an inverse covariance value greater or equal to 0.8.

iii) Unique Bayesian Networks and Prediction. In this paper we are exploring four different studies, each of which we want to explore the unique mechanisms, we consider each of the four studies as a study-cluster of one element and the related glasso-network (built earlier) as the consensus network for that study-cluster. Although consensus approaches are popular, here we are interested in exploring the study-specific mechanisms through that we call *unique-networks*. Given a generic graph $G = (V, E)$. We have m fixed graphs G_i such that $G_i = (V, E_i)$, where $V = 1, ..., n$ is the set of vertices(nodes) of the graph and $E_i = \{e_i\} = \{(u_{i1}, v_{i1}), ..., (u_{ik_i}, v_{ik_i})\}$, $k_i = \left|E_i\right|$ and $k_i \leq n(n-1)/2$. We define the unique function as $\Phi : G \mapsto G$, where, given $\hat{E}_i = \bigcup_{j=1, j\neq i}^{m} E_j$.

Definition 1: We define a function $\Phi(G_i)$ such that $\Phi(G_i) : (V, \{e_i : e_j \in E_i \text{ and } e_j \notin \hat{E}_i\})$
In other words, a *unique-network* contains only those edges present in no other condition-specific network. We choose to measure the reliability of the unique-networks through prediction using Bayesian Networks (BNs)[17,18] which naturally perform this using inference, given the graphical structure obtained using the genes involved in the unique-networks provided by *glasso*. Given the unique edges in the *glasso*-derived networks we first build one BN for each of the study-clusters using the R package *bnlearn*[19,20] and then identify the most predictive (how well it predicts other expression level values) and predictable (how well its expression level values are predicted) genes within (intra) and outside (inter) the study using the package *gRain*[21] and the leave one out cross validation technique. Given the m samples and n genes within each study we use m-1 samples as a training set and the remaining one as test set. Then, given the n-1 genes, we predict the expression value of the one left out. We compare the predicted with the real value, return 1 if they correspond and zero otherwise. We do this within all the studies and for all possible combinations of training and test sets of studies and genes. Finally, we average the amount of correctly-predicted values among the total predictions to obtain the correct-prediction for each gene. The idea is that genes that are predictive or predicted better within the selected study than on other studies are more likely to be relevant to the unique-network.

iv) Gene cards. As we detect study-specific sub-networks we also want to verify that our method captures study-specific genes. We query GeneCards encyclopaedia[1] to obtain the list of genes that are known to be involved in each cancer. We compare the list for each study to the others and select the genes that appear *only in the study under consideration*. To compare the unique-gene list for each type of cancer with the genes found in the corresponding unique-network, we apply a probability score developed in[6] used to test the significance of observing multiple genes with known function in a given cluster against the null hypothesis of this happening by chance. This score is based on the hypothesis that, if a given cluster, i of

size s_i, contains x genes from a defined functional group of size k_j, then the chance of this occurring by chance follows a binomial distribution and is defined by: $\Pr(\text{Observing} x \text{from group} j) = \binom{k_j}{x} p^x q^{k_j - x}$ where $p = \frac{s_j}{n}, q = 1 - p$ and n is the number of genes in the dataset. As in this paper, when k_j and x are very large Pr cannot be evaluated. Therefore we use the normal approximation of the binomial distribution where: $z = \frac{x-\mu}{\sigma}$, $\mu = k_j p$ and $\sigma = \sqrt{k_j pq}$. Values of z above zero mean that the probability of observing x elements from functional group j in cluster i by chance is very small (values of $z \geq 2.326$ correspond to a probability less than 1%). The test performed is the one tailed test.

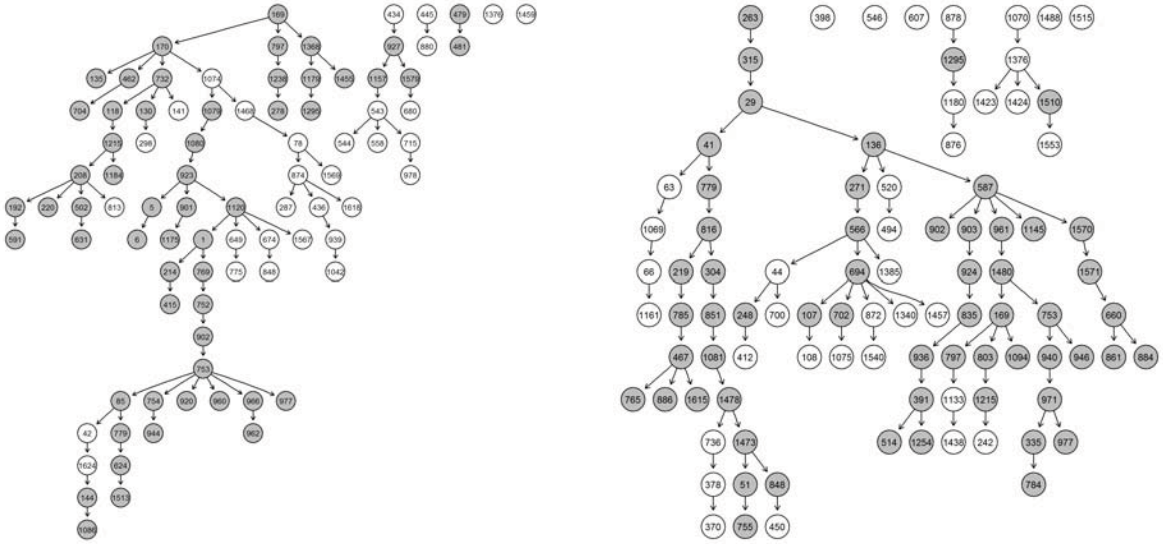
v) Logic and GUI. Finally a user interface has been developed using the R package *shiny* [22]. This interface allows the user to input the networks obtained with glasso and let the user choose which combination of unique networks to identify, using the logic operators AND and NOT. For example setting **1 AND 2 - NOT 3** will identify the sub-networks that study 1 and 2 have in common but do not appear in study 3. The unique sub-networks for that rule/pattern are identified and plotted on the interface together with the list of genes involved. Finally, the user has the possibility to save the network in a tiff file and the list of genes involved in csv format.

Results

In this study four cancer datasets are explored: breast, ovarian, medullary breast (a subtype of breast cancer) and lung, in human patients. Each dataset contains a different number of samples (see Table 1). The variable selection approach reduces the number of variables/genes to analyse from 54675 to 1629. Variable reduction is followed by the implementation of *glasso* with the parameter $\rho = 0.05$. Given the glasso networks for each study we consider only the edges that are present in the network under consideration but not in the others. Once the unique-edges are detected, the genes involved are used to build a BN for each study called unique-networks (U-Ns). An example of these networks is shown in Figure 1. The structure of the glasso U-Ns differ from the structure of the Bayesian U-Ns. In the Figures 1a and 1b the nodes with a grey background indicate genes with a predicted accuracy for the gene greater than 0.6 (based on our findings in [9]). Because of the study description in Table 1, we would expect breast cancer to be very similar (involving almost the same genes) to medullary breast cancer and slightly less similar to ovarian, but very different from lung cancer. This implies that the average internal prediction for each study will not differ much from the external prediction. The internal vs external prediction for each study shown in Figure 2 reveals, as expected a very clear difference only in Network 3 and 4, medullary-breast and lung cancer respectively, with a small difference in 1 and 3. This deduction is supported by the p-values obtained from the applied t-test as shown in Table 1. We now evaluate the significance of detecting the identified unique-genes by calculating the probability score using the normal approximation. For this paper s_i is the size of each unique network, k_j the number of genes in the unique gene-list obtained for each cancer type comparing the geneCards gene lists, x the number of genes that are present on both the unique network and the corresponding unique gene-list and n is the number of genes in the original unprocessed dataset. The results in Table 2 show the z-score and the corresponding p-value indicating that the probability of observing x elements from functional group j in cluster i by chance is in all four cases very small. This implies that the unique genes identified by our pipeline are highly significant in all studies.
Finally, Figure 3 shows the Logic Application interface. The example allows the user to visualize the unique sub-networks and the list of related genes that study 1 AND 4 have in common but do not appear in study 2.

Table 1: Cancer datasets description and t-test p-value

Study ID	Study title	Samples	t-test p-value
GSE18864	Triple Negative Breast Cancer	84	0.55
GSE9891	Ovarian Tumour	285	0.00
GSE21653	Medullary Breast Cancer	266	0.02
GSE10445	Adenocarcinoma and large cell Lung Carcinoma	72	0.00



(a) Bayesian U-N for medullary-breast cancer.

(b) Bayesian U-N for lung cancer.

Figure 1: Nodes with grey background indicate a prediction accuracy for the nodes greater than 0.6. Isolated nodes do not have connections due to the structure differences between glasso U-Ns and Bayesian U-Ns. Nodes are labelled with numbers (directly corresponding to the gene ID) for visualization purposes.

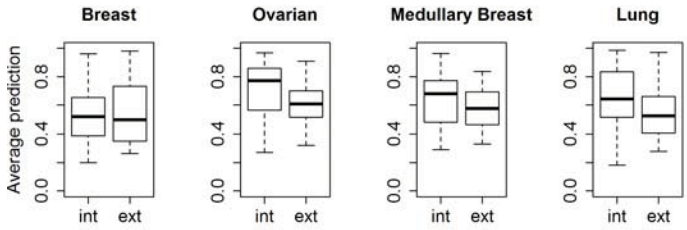


Figure 2: Internal vs External prediction accuracy for each study averaged among all genes involved in the related unique-network.

Table 2: Parameters values, z-score and p-value for each study.

Parameters values for each study						
Study ID	s_i	k_j	x	n	z-score	p-value
GSE18864	117	2982	11	54675	1.83	$\leq 3.4\%$
GSE9891	61	692	4	54675	3.68	$\leq 1\%$
GSE21653	89	0	0	54675	NaN	$\leq 1\%$
GSE10445	80	240	3	54675	4.47	$\leq 1\%$

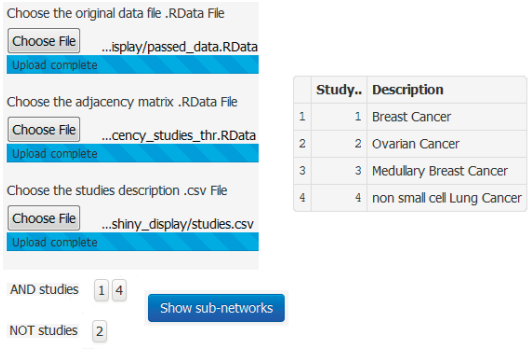


Figure 3: Logic Application interface.

Conclusions

We have developed a tool that aims to identify unique sub-networks and genes based upon a number of microarray studies. We explore networks and genes that are robust and unique to a pre-selected number of studies. We support our results using prediction accuracy and a score to test the significance of identifying a subset of unique genes. Furthermore, we created an application interface which allows the user to combine different studies through AND and OR logic operators. Based on the findings we conclude that our pipeline is a robust and reliable method to analyse large sets of transcriptomic data. It detects relationships between transcriptional expression of genes that are specific to different conditions and also highlights structures and nodes that could be potential targets for further research.

References

1. Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database. 2010;2010:baq020.
2. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics. 2003;19(suppl 1):i84–i90.
3. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–3297.
4. Steele E, Tucker A. Consensus and Meta-analysis regulatory networks for combining multiple microarray gene expression datasets. Journal of biomedical informatics. 2008;41(6):914–926.
5. Anvar SY, Tucker A, et al. The identification of informative genes from multiple datasets with increasing complexity. BMC bioinformatics. 2010;11(1):32.
6. Swift S, Tucker A, Vinciotti V, Martin N, Orengo C, Liu X, et al. Consensus clustering and functional interpretation of gene-expression data. Genome biology. 2004;5(11):R94.
7. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature genetics. 2003;34(2):166–176.
8. Zhang J, Lu K, Xiang Y, Islam M, Kotian S, Kais Z, et al. Weighted Frequent Gene Co-expression Network Mining to Identify Genes Involved in Genome Stability. PLoS Computational Biology. 2012;8(8):e1002656.
9. Bo V, Curtis T, Lysenko A, Saqi M, Swift S, Tucker A. Discovering Study-Specific Gene Regulatory Networks. PLoS ONE. 2014;9(9):e106524.
10. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research. 2002;30(1):207–210.
11. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–315.
12. Lu J, Bushel PR. pvac: PCA-based gene filtering for Affymetrix arrays; 2010. R package version 1.12.0.
13. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901;2(11):559–572.
14. Friedman J, Hastie T, Tibshirani R. glasso: Graphical lasso- estimation of Gaussian graphical models; 2014. R package version 1.8.
15. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–441.
16. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics. 2006;34(3):1436–1462.
17. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: The combination of knowledge and statistical data. Machine learning. 1995;20(3):197–243.
18. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. Journal of computational biology. 2000;7(3-4):601–620.
19. Scutari M. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:09083817. 2009;.
20. Scutari M, Scutari MM. Package bnlearn. 2012;.
21. Højsgaard S. Graphical Independence Networks with the gRain package for R. Journal; 2012.
22. RStudio, Inc . shiny: Web Application Framework for R; 2014. R package version 0.9.1.

Leveraging an Electronic Health Record-Linked Biorepository to Generate a Metformin Pharmacogenomics Hypothesis

Matthew K. Breitenstein, PhD[1,2], Liewei Wang, MD, PhD[1], Gyorgy Simon, PhD[2], Euijung Ryu, PhD[1], Sebastian M. Armasu, MS[1], Balmiki Ray, MBBS[1], Richard M. Weinshilboum, MD[1], Jyotishman Pathak, PhD[1]
[1]Mayo Clinic, Rochester, MN; [2]University of Minnesota, Minneapolis, MN

Abstract

Metformin is a first-line antihyperglycemic agent commonly prescribed in type 2 diabetes mellitus (T2DM), but whose pharmacogenomics are not clearly understood. Further, due to accumulating evidence highlighting the potential for metformin in cancer prevention and treatment efforts it is imperative to understand molecular mechanisms of metformin. In this electronic health record(EHR)-based study we explore the potential association of the flavin-containing monooxygenase(FMO)-5 gene, a biologically plausible biotransformer of metformin, and modifying glycemic response to metformin treatment. Using a cohort of 258 T2DM patients who had new metformin exposure, existing genetic data, and longitudinal electronic health records, we compared genetic variation within FMO5 to change in glycemic response. Gene-level and SNP-level analysis identified marginally significant associations for FMO5 variation, representing an EHR-driven pharmacogenetics hypothesis for a potential novel mechanism for metformin biotransformation. However, functional validation of this EHR-based hypothesis is necessary to ascertain its clinical and biological significance.

Introduction

Metformin is a first-line antihyperglycemic agent commonly prescribed for type 2 diabetes mellitus (T2DM) patients[1], whose pharmacogenomics are not clearly understood[2], but are thought to be absent of biotransformation[3]. Further, glycemic response to metformin is variable[3] and serious adverse reactions to metformin have been known to occur[4]. Due to increasing evidence highlighting the potential for metformin in cancer prevention and treatment, it is imperative to understand molecular mechanisms of metformin further.

Background

Metformin is primarily utilized to regain glycemic control in diabetic or pre-diabetic patients. Metformin is a relatively safe antidiabetic therapy[5]. However, serious adverse reactions can occur[4] and there is considerable variation in glycemic response to metformin, with ~30% of patients unable to achieve glycemic control with metformin[3]. While genetic factors may partially explain clinical glycemic response to metformin due to pharmacokinetic(PK) determinants[3], the transportation throughout the body variation, the identification and impact of metformin pharmacodynamic(PD) determinants, the physiological and biochemical impact of metformin in the body, remains uncertain[2]. Regarding PKs, Metformin is thought to not be metabolized[3], with absorption of metformin known to occur in the small and large intestines[5]. Uptake of metformin from the blood is known to occur in the kidneys and liver[2], but can be reasonably assumed to occur in any tissue with abundance of organic cation transporters (OCT). Eventually metformin is excreted unchanged in the urine[5]. Regarding PDs, metformin works primarily by inhibiting hepatic glucose production by reducing gluconeogenesis in the liver[6] and is also known to reduce intestinal glucose absorption[7]. Further, metformin appears to improve glucose uptake and utilization systemically[3].

Metformin is a nitrogen-rich biguanide. Flavin-containing monooxygenases(FMO)-5 has demonstrated narrow substrate specificity, but has been known to catalyze oxygenation of nitrogen-containing drugs[8]. FMO5 is expressed in the kidneys and liver[8]. The FMO5 gene exists near PRKAB2, a known PD regulator of metformin response, away from the single gene cluster for the remaining FMOs in chromosome 1q23-q25 region. Metformin is excreted unchanged in the urine[5], hinting that metformin does not undergo biotransformation. However, studies such as these do not produce 100% yield, hinting at room for deviation from this paradigm. While metformin is thought to be absent of biotransformation[3], it is biologically plausible that FMO5 might carry out N-oxygenation of metformin.

FMOs show overlapping substrate specificity among family members[8]; a signal corresponding to FMO5 might also correspond to an additional FMO gene. All FMOs contain eight coding exons that share 50 to 80% sequence identity, with mutant FMOs are known to react to alternative chemical sites[9]. FMOs are localized in the endoplasmic

reticulum of the cell whose expression is tissue-specific[8]. The extent of which reactions are catalyzed by FMOs in vivo cannot be determined by measuring end products excreted in bile or urine[10].

The primary purpose of this study was to add clarity to metformin pharmacogenomics by understanding the impact of common variants in the FMO5 gene on altered glycemic response in a clinical population derived from an EHR-linked biorepository. Due to some shared functional similarity among genes in the FMO gene family, we selected the remaining FMO genes (FMO1 – FMO4) as exploratory gene candidates as our secondary hypothesis.

Methods

In this EHR-linked genetic study, both the approaches for obtaining clinical phenotypes and genotypes had important considerations for both study design and study interpretation. Our primary hypothesis of interest holds that genetic variation within FMO5 has potential to modify glycemic response to metformin monotherapy. Secondary to the primary hypothesis is an exploratory hypothesis that posits similar potential associations for FMO1 – FMO4 due to functional similarity[8]. However, their function is not identical. Further, due to the close proximity of the FMO1 – FMO4 to each other and their relative distance from FMO5 on chromosome 1q21 our secondary hypothesis is considerably weaker than our primary hypothesis for FMO5. In this study, we utilized the longitudinal EHR at Mayo Clinic and genome-wide association study (GWAS) data from the subjects enrolled in the Mayo Genome Consortia[11].

Clinical Phenotypes

The application of EHR-based phenotypes dramatically impacts study design and interpretability of findings. In this study we had 4 key phenotype aspects to consider: 1) T2DM phenotype, 2) metformin exposure phenotype, and 3) change in A1c. First, attribution of a T2DM phenotype was performed using a modified methodology developed by eMERGE[12]. A key point of differentiation is that our T2DM phenotype relied on diagnosis codes and did not initially consider laboratory values or medication. However, our second and third considerations relied on lab values and medication exposure events that were more specific than the criteria for the eMERGE T2DM phenotype algorithm. Second, our metformin exposure period was designated as a new prescription of metformin that extended ≥6 months to ensure adequate primary care visits, multiple A1c measures, and maintenance dose achievement. Since our study aimed to understand genomic variation in relation to patients who respond or do not respond to metformin, maintenance dose was not a consideration. To accurately populate this metformin exposure phenotype our study design required longitudinal data access from primary care patients. Specifically, study inclusion criteria required ≥1 year of patient history and ≥2 primary care visits to ensure accurate capture of the first date of metformin exposure, which aimed to exclude patients that were false positives for a new recorded exposure to metformin due to medication reconciliation that occurred at transfer of primary care. Metformin exposure events were ascertained using a combination of validated structured and semi-structured EHR data collection methodologies that leveraged our prior work[13,14] where a total of 1 generic name (metformin) and 4 brand name medications (Fortamet®, Glucophage®, Glumetza®, and Riomet®) were queried. Patients with <6 months of metformin exposure or on combination drugs that included metformin or other prescribed antidiabetic drugs during the ≥6 month exposure period were excluded from the study. Third, to compare the association of genetic modification to glycemic response to metformin, measures of A1c were compared prior to metformin exposure and during the period of metformin exposure following a 6-month period of delay to allow for the achievement of maintenance dosage. A1c measures were required ≤6 months prior to metformin exposure and ≥6 months after metformin exposure. A1c measures were averaged across sections that occurred before and up to the date of metformin exposure. A1c measures were averaged across the period occurring ≥6 months after initial metformin exposure and until either metformin exposure ceased or anti-diabetic combination therapy

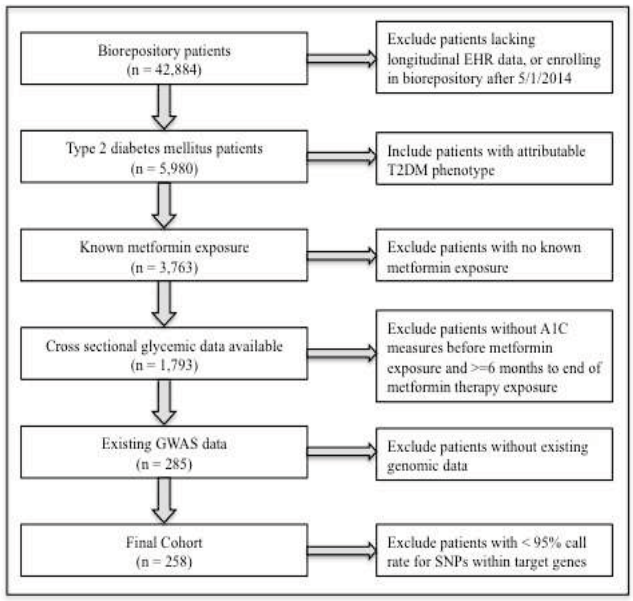


Figure 1: Study Cohort Development Process

was initiated. This approach minimizes the impact of any one A1c measure and biases change in A1c measures towards the null.

Genotyping and Quality Control

The MayoGC stores existing GWAS data generated from multiple studies. These data were harmonized to the forward strand mapped to become on the same strand as the 1000 genome cosmopolitan reference population. Genotypes for unmappable or ambiguous SNPs were excluded. We selected SNPs 20 kb upstream and downstream of each gene using 1000 genomes project variants and NCBI build 37 as the reference genome. By this mapping rule, a total of 1,381 SNPs were mapped to the 5 genes, but only 205 SNPs were available in the genotype data. Further, due to their proximity the FMO1, FMO2, FMO3, and FMO4 genes some SNPs belong to multiple genes. For the remaining SNPs, two main quality control filters were applied:(i) SNPs with unacceptable high rates of missing genotype calls (>10%); and (ii) monomorphic SNPs were excluded. The quality control of the genotype data was performed by PLINK v1.07[15]. A detailed diagram of cohort development is found in **Figure 1.**

Analysis

The SNP-level and gene-level analyses were performed on the final analysis cohort where 258 Caucasian subjects had metformin exposure, complete EHR data, and 90 SNPs after quality control. In the analysis, we adjusted for age, gender, and morbid obesity (BMI ≥35), a known modifier of T2DM state[1] as fixed covariates in our model. Age and BMI measures were calculated at first recorded exposure to metformin. The endpoint of change in A1c was transformed using Van der Waerden rank, otherwise known as rank based inverse Gaussian, to normalize and accommodate linear regression modeling. Batch adjustment did not change the results of GWAS data (data not shown) and was not adjusted in the displayed results. SNP-level and gene-level results were described, but not displayed, after application of Bonferroni correction.

Table 1: Demographics (n=258)

Variable	**n (%)**
Female, N(%)	89 (34.5)
Male, N (%)	169 (64.5)
BMI <30, N (%)	64 (24.8)
BMI(≥30 to <35 kg/m^2)	100 (38.8)
BMI ≥35 (kg/m^2)	93 (36.1)
Median A1c >7.0 (DCCT %), N (%)	101 (39.1)
Change in A1c (DCCT %), median (range)	0.07 (-6.45, 3.51)
Age (years), median (range)	64 (30, 84)

SNP-Level Analysis

SNP-level analyses were performed on each SNP in FMO genes pertaining to both our primary and secondary hypothesis to identify top SNPs and determine directionality of their associations. Using Van der Waerden rank transformation on change in A1c, linear regression models were applied adjusting for age, gender and morbid obesity. Coefficient estimates were calculated per minor allele, that is, with each minor allele, the A1c level changes by 'beta'. SNP-level results are displayed as unadjusted for multiple testing. Finally, conditional analysis was performed to identify potentially independent SNPs in each gene. Locus Zoom plots were also created for better visualization using the LD in the 1000 Genomes European reference population from March 2012 release.

Gene-Level Analysis

Gene-level tests were performed using principal component analysis (PCA)[16]. For each gene, principal components (PC) were created using linear combinations of ordinally scaled SNPs (i.e., 0, 1, 2 copies of minor allele) and the smallest set of resulting principal components that explained at least 90% of the SNP variance within the gene was included in linear regression models. Instead of including the entire set of SNPs for each gene, the PC approach reduces the degrees of freedom, avoids model fitting issues due to multi-collinearity of the SNPs from linkage disequilibrium (LD) and potentially improves the statistical power. Finally, we computed the likelihood ratio test (LRT) to assess overall significance of a gene by comparing the null model containing only the covariates with the full model containing covariates and the set of resulting principal components. The statistical package R 2.15.0 was utilized for the gene-level analysis. Plots of LD displaying r^2 for FMO5 gene was created using Haploview v 4.2.

Results

Our EHR-based phenotyping algorithm identified 1,793 T2DM subjects **(Figure 1)**. Among those, 258 subjects had 90 SNP data that passed quality control criteria. Cohort demographics can be found in **Table 1.** The estimates for male (Coefficient=0.0435, P-value=0.737), age (Coefficient=-0.0009, P-value=0.881), morbid obesity (Coefficient=0.2214, P-value=0.083) were not significantly associated with change in A1c at alpha=0.05 significance level in the univariate analysis. Further, none of the covariates were associated with change in A1c at alpha=0.05 significance level in a multivariate model.

SNP-Level Results

Of the 5 candidate genes, only FMO5 had SNPs that demonstrated a potentially significant association **(Table 2)**. After adjusting for multiple testing rs7541245, the top SNP in FMO5, was marginally significant, but since this signal is very close to passing correction (0.00188-observed vs. 0.00161-Bonferonni threshold) it was deemed appropriate for consideration. None of the SNPs in FMO1-FMO4 gene cluster were found to be significant. Among 31 genotyped SNPs within FMO5 gene **(Figure 2)**, 4 SNPs had p-values less than 0.05 for the association with a decrease in glycemic response during metformin exposure, with rs7541245 having the most significant signal. The FMO5 linkage disequilibrium (LD) plot (not shown due to space constraints) contained 4 LD blocks and appeared to show 9 independent SNPs. The conditional analysis that adjusted for the top most significant SNP in each gene and clinical covariates was performed. FMO5 rs7541245 was the main signal on FMO5 gene as no SNPs reached p-values less than 0.05 which pointed to the remaining SNPs within FMO5 being in high LD with rs7541245 and hence, not independent.

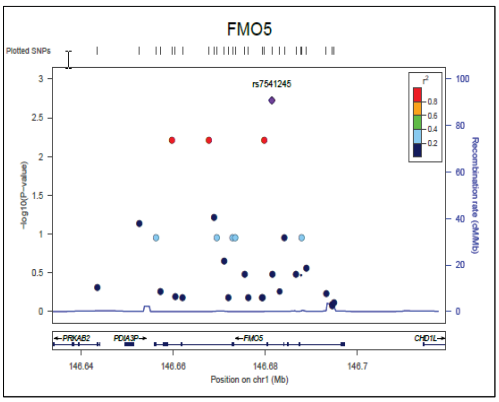


Figure 2: Locus Zoom Plot for FMO5 association with glycemic response to metformin

Table 2: SNP-Level Analysis[+]

Gene Name	Genotyped SNPs (n)	Top SNP	Minor Allele	Major Allele	MAF	BETA	95% CIs	P-value
FMO5	31	rs7541245	A	C	0.0311	-0.7885	(-1.28;-0.297)	**0.00188***
FMO4	15	rs2076322	G	A	0.1395	0.2061	(-0.055;0.467)	0.12270
FMO3	19	rs1920145	C	T	0.3346	-0.1076	(-0.296;0.081)	0.26530
FMO2	14	rs12752688	T	C	0.1434	0.2256	(-0.025;0.476)	0.07885
FMO1	12	rs13376631	G	A	0.1376	0.1997	(-0.062;0.461)	0.13560

MAF = minor allele frequency, * marginally significant after correction for multiple testing
[+]Only top SNPs displayed due to manuscript space constraints

Gene-Level Results

Our primary hypothesis for the FMO5 gene, represented by 5 PCs and 31 genotyped SNPs, was marginally significantly associated (p=0.0185) with glycemic response **(Table 3)** after controlling for age, gender and morbid obesity. No significant associations were identified for our secondary hypothesis tests of the remaining FMO genes.

Table 3: Principal Component Analysis

Gene	Genotyped SNPs (n)	nPCs	P-value
FMO5	31	5	**0.0185**
FMO4	12	4	0.5623
FMO3	14	5	0.5464
FMO2	19	4	0.3581
FMO1	15	6	0.5479

Discussion

In this study, we leverage EHR-linked biorepository data and EHR-based phenotyping methods to study common variants within FMO5, our gene of primary interest. While the FMO5 gene appeared to be of marginal significance in relation to glycemic response to metformin, our secondary hypothesis for the remaining FMO genes demonstrated no significance. Given the study design and execution of phenotypes, results of this study can be interpreted most accurately as pharmacogenetics hypothesis generating. However, this hypothesis could represent a novel mechanism for the biotransformation of metformin or other potential mechanism of metformin action that has been previously unidentified. Additional studies are needed; functional studies are potentially warranted.

In our study not all SNPs within candidate genes were available for analysis due to GWAS genotyping being originally performed for other studies. No effect difference was observed between cohort batches which hint that our findings were not biased due to original patient selection criteria or genotyping criteria, however potential for heterogeneity remains. Having all patients with T2DM and metformin allowed for us to identify genetic variation as the consideration of interest. However, the limited sample size paired with a relatively weak clinical outcome had potential to bias associations towards the null. While utilizing a clinical endpoint enabled us to engage in exploratory research, our signal strength was limited by modest cohort size (n=258) and the study criteria design. Specifically, by removing patients with <6 months of metformin exposure during metformin exposure we potentially removed patients who were complete non-responders to metformin or who experienced an adverse

reactions to metformin. By study design these would not have been able to attain glycemic control with metformin, biasing our outcome phenotype towards positive glycemic response (i.e. decreased A1c) to metformin.

Alterations in FMO genes are known to induce differential biotransformation of nitrogen-rich compounds, such as metformin[10]. In this study, it appeared that the utility of metformin (i.e. glycemic response) is impaired by alterations in the FMO5 gene, hinting that potential biotransformation of metformin might be occurring in the normal FMO5 gene product. Our finding hints that metformin conjugates resulting from metformin biotransformation via FMO5 might be responsible for the anti-diabetic effects of metformin. Should these findings be confirmed by functional studies, this hypothesis could represent a novel mechanism for the biotransformation of metformin and mechanism of metformin action that has been previously unidentified.

Conclusion

FMO5 appears to be marginally significantly associated with decreases in glycemic response after exposure to metformin, representing an EHR-driven pharmacogenetics hypothesis that could represent a novel mechanism for the biotransformation of metformin that has been previously unidentified. Functional validation of this hypothesis is warranted to ascertain its clinical and biological significance.

Acknowledgments

This work has been supported by NIH grant U19-GM61388-13, PGRN Network Resource, Pharmacogenomics of Phase II Drug Metabolizing Enzymes.

References

1. Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.
2. Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.
3. Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.
4. Bailey CJ, Path MRC, Turner RC. Metformin. The New England Journal of Medicine. 1996;334(9):574-579.
5. Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.
6. Hundal RS, Krssak M, Dufour S, et al. Mechanism by Which Metformin Reduces Glucose Production in Type 2 Diabetes. Diabetes. 2000;49.
7. Sakar Y, Meddah B, Faouzi MYA, Cherrah Y, Bado A, Ducroc R. Metformin-Induced Regulation of Intestinal D-Glucose Transporters. Journal of Physiology and Pharmacology. 2010;61(3):301-307.
8. Lattard V, Zhang J, Cashman JR. Alternative Processing Events in Human FMO Genes. Molecular Pharmacology. 2004;65(6):1517-1525.
9. Joosten V, van Berkel WJH. Flavoenzymes. Current Opinion in Chemical Biology. 2007;11:195-202.
10. Ziegler DM. Flavin-Containing Monooxygenases: Catalytic Mechanism and Substrate Specificities. Drug Metabolism Reviews. 1988;19(1):1-33.
11. Bielinski SJ, Chai HS, Pathak J, et al. Mayo Genome Consortia: A Genotype-Phenotype Resource for Genome-Wide Association Studies With an Application to the Analysis of Circulating Bilirubin Levels. Mayo Clinic proceedings. 2011;86(7):606-614.
12. Kho AN, Hayes MG, Rasmussen-Torvik L, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association : JAMIA. Mar-Apr 2012;19(2):212-218.
13. Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. JAMIA. Mar-Apr 2010;17(2):131-135.
14. Pathak J, Murphy SP, Willaert BN, et al. Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. American Medical Informatics Association Annual Symposium. 2011:1089-1098.
15. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics. Sep 2007;81(3):559-575.
16. Gauderman WJ, Murcray C, Gilliland F, Conti DV. Testing association between disease and multiple SNPs in a candidate gene. Genetic epidemiology. Jul 2007;31(5):383-395.

Novel Application of Junction Trees to the Interpretation of Epigenetic Differences among Lung Cancer Subtypes

Arturo Lopez Pineda, MS[1], and Vanathi Gopalakrishnan, PhD[1]
[1]The PRoBE Lab, Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, PA

Abstract

In this era of precision medicine, understanding the epigenetic differences in lung cancer subtypes could lead to personalized therapies by possibly reversing these alterations. Traditional methods for analyzing microarray data rely on the use of known pathways. We propose a novel workflow, called Junction trees to Knowledge (J2K) framework, for creating interpretable graphical representations that can be derived directly from *in silico* analysis of microarray data. Our workflow has three steps, preprocessing (discretization and feature selection), construction of a Bayesian network and, its subsequent transformation into a Junction tree. We used data from the Cancer Genome Atlas to perform preliminary analyses of this J2K framework. We found relevant cliques of methylated sites that are junctions of the network along with potential methylation biomarkers in the lung cancer pathogenesis.

Introduction and Background

Lung cancer is the leading cause of human cancer death in the United States, with estimated yearly casualties of over 160,000 [1]. Among all lung cancers the most frequent subtypes are adenocarcinoma (ADC) and squamous cell carcinoma (SCC), accounting for 38.5% and 20% of all cases respectively [2]. Identifying molecular differences between these two subtypes is important to enable clinicians to select patients who will likely benefit from a given drug regimen, and also in selecting those patients who will avoid toxicity from the treatment [3]. It has been suggested that the ADC and SCC develop through distinct pathogenetic pathways, resulting in epigenetic alterations [4].

DNA methylation is an epigenetic alteration that creates molecular changes to the environment of the DNA. This alteration occurs when a methyl group is attached to a specific location of the DNA, typically in sites where the sequence cytosine-phosphate-guanine (CpG) is abundant. This epigenetic alteration has the effect of silencing gene transcription, potentially removing important functions in the protein pathways. Recent studies have found that DNA methylation can have an effect on the progression [5], and recurrence [6], of the cancer into a more aggressive form.

Distinct DNA methylation signatures between ADC and SCC have been found in studies targeting candidate genes in lung cancer [7]. However, there is still a need for understanding the mechanisms of DNA methylation for future epigenetic therapies, such as reversal of DNA methylation. This therapy has shown promising results using a technique called active demethylation that promotes DNA repair [8].

Traditional methods for analysis of DNA methylation microarray technology often investigate open research questions such as: How do differentially methylated sites interact among one another? Is there a network topology that might help discover differences between lung cancer subtypes? Balbin et al. [9] recently proposed a framework for the reconstruction of gene network topology in lung cancer combining multiple 'omic' data and then identifying functional networks associated with them. This framework relies on the identification of differentially expressed pathways from the 'omic' data. To achieve this task they use SPIA [10], an algorithm that combines the evidence of enrichment analysis (log-fold change differential expression) from the 'omic' data and the perturbation it has on the known pathways in KEGG [11]. Martini et al. [12] proposed an algorithm for the reconstruction of relevant network topologies. They start with known pathways found in KEGG which they transform into Junction trees[13] for human interpretation. Pradhan et al. [14] used microarray data to select the differentially expressed genes and then used the known interactions in the BioGrid [15] platform to reconstruct the network topology.

This paper describes a novel workflow, called Junction trees to Knowledge (J2K), for performing *in silico* analysis of DNA methylation datasets. It addresses the post-classification problem of characterizing biomarkers that are responsible for disease classification at the sub-clinical level. It makes use of Bayesian networks (BNs) [16], which traditionally have been used in other domains to perform probabilistic inference; and Junction trees (JTs) [13], which have been vastly applied to propagate belief over a network and compute exact posterior probabilities [17].

While there are many computational algorithms that can assist in the creation of JTs, their application to modeling 'omic' data is relatively new. To our knowledge, the use of junction tree representation to simplify the BN has not been explored adequately with biomedical data. While the literature supports their use for efficient identification of proteins from tandem mass spectra [18], it is unclear whether a representation that is created for purely computational efficiency can also provide biologically relevant results. We believe that our research will help us understand this aspect better.

Materials and Methods

Our workflow, as shown in Figure 1, first discretizes the features in the data using MDLPC [19], and selects those that best distinguish the target class via feature selection with the ReliefF [20] algorithm. Then it builds a BN using EBMC [21], and finally it transforms the directed network into a JT [13]. The remaining parts of this section describe these algorithms, including the in-house developed JT creation algorithms.

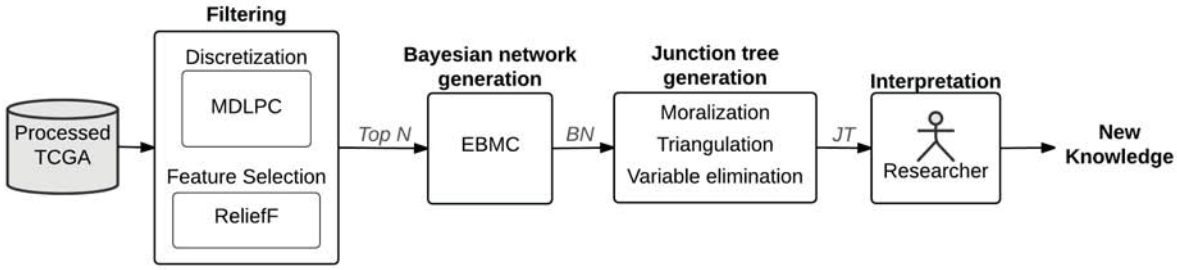


Figure 1. Empirical workflow of TCGA data to directed graph (BN) to undirected graph (JT) to Knowledge (J2K)

Dataset

The Cancer Genome Atlas (TCGA) is a public repository of genomic data supervised by the National Cancer Institute (NCI) that aims to characterize human cancers. We extracted DNA methylation intensity profiles for 197 tumor samples (65 ADC and 132 SCC) from the TCGA data portal for lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC, [22]). The microarray platform used for analysis was the Illumina® Infinium HumanMethylation 27k (27,578 variables representing methylation sites).

Data Preprocessing

Discretization. The methylation intensity for each methylation site in the microarray data is a continuous value. We partitioned this value into intervals using the Fayyad and Irani's Minimum Description Length Principle Cut (MDLPC) [19]. This algorithm selects a cut point that minimizes the joint entropy of the resulting subintervals. It then continues to partition recursively until no cut point can be selected. The resulting set of intervals constitutes the values for that methylation site. Features refer to variable-value pairs.

Feature Selection. A subset of the most informative features from the microarray data were selected using the ReliefF feature selection algorithm [20]. This is a multivariate filter that sequentially evaluates every instance to estimate how well a feature can distinguish the target class given the instances in the same neighborhood. In the end, the top scoring features are retrieved. The ordering from this selection is also used in subsequent elements of our workflow.

Bayesian Networks

A Bayesian network [16] is a probabilistic graphical model that explains a given set of discrete data. It comprises a set of nodes (methylation sites) that represent random variables and a set of arcs among the nodes that represent probabilistic dependence. The posterior probability of an event (disease) D occurring, given that an event (symptom) S is observed is calculated using the well-known Bayes' formula [23].

The *Efficient Bayesian Multivariate Classification (EBMC)* [21] is a recent method for learning a BN from data. It uses a greedy search to find a constrained BN that best predicts a target node, similar to the Bayesian Rule Learning (BRL) algorithm [24]. It initially starts with an empty model and then it identifies a set of nodes that are parents of the target and predicts it well. EBMC then transforms the temporary network structure into a statistically equivalent one (Augmented Naïve Bayes or ANB [25]) where the parents of the target become children of the target with arcs

among them. It then iterates the whole process until no set of parents can be added to the target node to improve the prequential score [26]. EBMC has been shown to perform well at binary outcome prediction using high-dimensional discrete datasets [27]. This method differs from other BN-learning algorithms in at least two ways: (a) no ordering of features is required and (b) the representation of the learned BN is an augmented Naïve Bayes structure.

Junction Trees

A Junction tree [13] is a tree-structured undirected graph, whose nodes correspond to cliques of features (or methylated sites in our case), and whose links connect pairs of cliques that have features in common. A clique is a subset of nodes in an undirected graph where any two nodes are connected by an edge. In order to create a JT, three steps are needed:

1. Moralization. Starting from a directed graph, such as a BN, the directionality of the edge is removed by connecting or 'marrying' the set of nodes that share common children but do not have direct edges between them. This yields an undirected moral graph.
2. Triangulation. In the undirected moral graph, all cycles containing four or more nodes must be triangulated. This process involves iteratively adding extra edges to eliminate such cycles of four or more nodes (chord-less cycles). There are an exponential number of triangulation possibilities, depending on the number of nodes involved in the cycle, which have been solved by using a predefined ordering of nodes to be triangulated. Triangulation is an NP-hard problem. We order the nodes for this triangulation process based on the ReliefF scoring to make the algorithm efficient. The triangulated graph containing cliques is used for node elimination, which creates a JT.
3. Node Elimination. A new tree-structured undirected graph (empty JT) is first constructed by following the node elimination algorithm. Then, a node is selected for elimination, and its containing clique is added to the JT. Next, the node and its incident edges are eliminated from the triangulated graph and the process is repeated until no other nodes are available. Node elimination is also an NP-hard problem. Hence, a predefined ordering of nodes has to be used. The JT must satisfy a property called the 'junction property' (*running intersection property*), meaning that if a feature is contained in two cliques, then it must also be contained in every clique on the path that connects them. The order in which the nodes were eliminated is based on the ReliefF scoring.

Results and Discussion

The MDLPC [13] algorithm transforms all continuous methylation profiles into discrete binned data containing at least two bins for each methylation site. Since MDLPC is a supervised discretization method, it also has the side effect of discarding those features with only one bin. For the datasets that we used, the algorithm selected 7,908 features out of 27,579.

The ReliefF algorithm [20] ranks all features according to their impact in differentiating between the classes. To demonstrate proof of concept, we selected the top ranked 30 methylated sites to create BNs. The area under the receiver operating characteristic curve (AUC) when evaluating these BNs over a stratified 10-fold cross-validation is 0.988. This level of classification performance is reasonable for this problem, because the two subtypes of lung cancer are known to be fairly distinct.

The structure of the BN created using the EBMC algorithms applied to the data is shown in Figure 2. As expected, there are multiple connections between the target node (class node) and the rest of the selected nodes. There are two fairly distinct clusters of 10 interconnected methylated sites (see Figure 2). The network topology produced by EBMC is an augmented naïve Bayes (ANB) structure. The corresponding gene IDs to where those methylation sites are located is used after this step (i.e. methylation site cg18515587 is located in gene SELENBP1).

As a conceptual method, we proposed the use of JTs as way of clarifying the structure of BNs. In Figure 3, we show the corresponding JT representation of the BNs seen in Figure 2. In the EBMC-derived JT representation we can start to think of new hypothesis with a greater biological relevance. For example, looking at the central clique (central circle) it is easy to see that there are key molecules that are worthwhile examining carefully, because a perturbation in this clique would have an impact on the entire structure. The central clique has the following genes: SELENBP1, AKR1B10, LPP, SH3BP1, BCDIN3, KRTHB4, MAPK8, TCF2. We used the suite NextBio® to test the association of this clique to different tissues and diseases in the known literature and curated studies. This suite finds that the central clique is associated with the epithelial cells of nasal turbinates, and the epithelial cells of bronchial large airways, and that it also correlates with esophageal cancer cell line OE21. The clique (and specially SELENBP1) is associated to the Selenium binding protein which is considered to be a tumor suppressor and a prognostic marker [28].

The BN shown in this manuscript is a Bayesian network classifier (BNC) created using the EBMC learning algorithm. The resulting structure of this BN improves the classification of the target node but does not to capture all the probabilistic relationships between features. We plan to use other BN learning algorithms to further explore this novel research area.

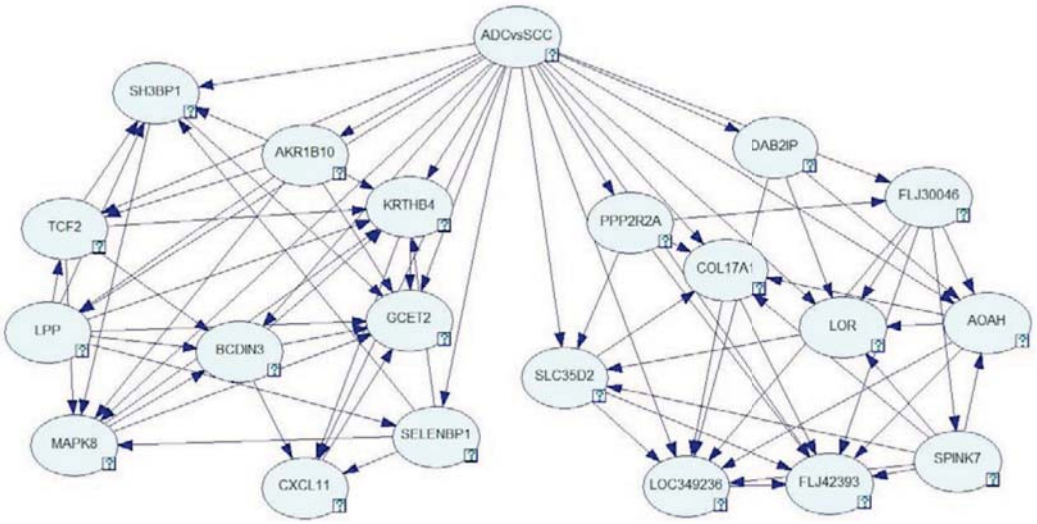


Figure 2. EBMC-generated BN model for the classification task ADC_{tumor} vs SCC_{tumor}.

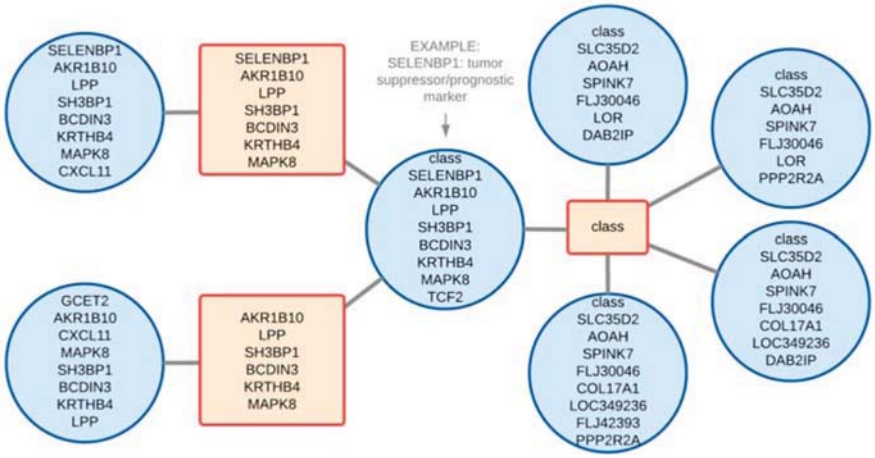


Figure 3. EBMC-derived JT, where the squares represent junctions while the circles represent cliques. An example is provided to show the importance of the JT to identify central cliques with important genes.

Conclusion

In this research, we have tested a novel concept called J2K framework to transform biological data from directed to undirected graphs via the application of JT generation algorithms. We applied a series of algorithms for transforming epigenomic data of lung cancer into a graphical representation that is interpretable for human researchers. Our study can easily be generalized into other types of epigenomic and genomic data, and we plan on testing it with other biomedical datasets. Particularly, we have found cliques of methylated sites that are of interest for the differentiation of lung cancer subtypes.

Acknowledgments: We thank Dr. Gregory F. Cooper for providing the EBMC java code.

Grant support: The research reported in this publication was supported by the following grants from the National Institutes of Health: National Cancer Institute Award Number P50CA90440, National Library of Medicine Award Number R01LM010950, and National Institute of General Medical Sciences Award Number R01GM100387. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1 Cancer statistics, 2012. 2012;**62**:10–29. doi:10.3322/caac.20138

2 Cruz Dela CS, Tanoue LT, Matthay RA. Lung Cancer: Epidemiology, Etiology, and Prevention. *Clinics in Chest Medicine* 2011;**32**:605–44. doi:10.1016/j.ccm.2011.09.001

3 Langer CJ, Besse B, Gualberto A, *et al.* The evolving role of histology in the management of advanced non-small-cell lung cancer. *J Clin Oncol* 2010;**28**:5311–20. doi:10.1200/JCO.2010.28.8126

4 Lockwood WW, Wilson IM, Coe BP, *et al.* Divergent genomic and epigenomic landscapes of lung cancer subtypes underscore the selection of different oncogenic pathways during tumor development. *PLoS One* 2012;**7**:e37775–5. doi:10.1371/journal.pone.0037775

5 Towle R, Truong D, Hogg K, *et al.* Global analysis of DNA methylation changes during progression of oral cancer. *Oral Oncol* 2013;**49**:1033–42. doi:10.1016/j.oraloncology.2013.08.005

6 Sato T, Arai E, Kohno T, *et al.* DNA methylation profiles at precancerous stages associated with recurrence of lung adenocarcinoma. *PLoS One* 2013;**8**:e59444–4. doi:10.1371/journal.pone.0059444

7 Rauch TA, Wang Z, Wu X, *et al.* DNA methylation biomarkers for lung cancer. *Tumor Biol* 2012;**33**:287–96. doi:10.1007/s13277-011-0282-2

8 Barreto G, Schäfer A, Marhold J, *et al.* Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. *Nature* 2007;**445**:671–5. doi:10.1038/nature05515

9 Balbin OA, Prensner JR, Sahu A, *et al.* Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat Commun* 2013;**4**:2617–7. doi:10.1038/ncomms3617

10 Tarca AL, Draghici S, Khatri P, *et al.* A novel signaling pathway impact analysis. *Bioinformatics* 2009;**25**:75–82. doi:10.1093/bioinformatics/btn577

11 Ogata H, Goto S, Sato K, *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;**27**:29–34. doi:10.1093/nar/27.1.29

12 Martini P, Sales G, Massa MS, *et al.* Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res* 2013;**41**:e19–9. doi:10.1093/nar/gks866

13 Lauritzen SL, Spiegelhalter DJ. *Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems*. 1988. doi:10.2307/2345762

14 Pradhan MP, Desai A, Palakal MJ. Systems biology approach to stage-wise characterization of epigenetic genes in lung adenocarcinoma. *BMC Syst Biol* 2013;**7**:141–1. doi:10.1186/1752-0509-7-141

15 Stark C, Breitkreutz B-J, Reguly T, *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9. doi:10.1093/nar/gkj109

16 Neapolitan RE. *Probabilistic Reasoning in Expert Systems*. 2012.

17 Serang O. The probabilistic convolution tree: efficient exact Bayesian inference for faster LC-MS/MS protein inference. *PLoS One* 2014;**9**:e91507–7. doi:10.1371/journal.pone.0091507

18 Serang O, Noble WS. Faster Mass Spectrometry-Based Protein Inference: Junction Trees Are More Efficient than Sampling and Marginalization by Enumeration. *IEEE/ACM Trans Comput Biol and Bioinf*;**9**:809–17. doi:10.1109/TCBB.2012.26

19 Fayyad U, Irani K. BEACON eSpace at Jet Propulsion Laboratory: Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. 1993.

20 Kononenko I, Šimec E, Robnik-Šikonja M. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. *Applied Intelligence* 1997;**7**:39–55. doi:10.1023/A:1008280620621

21 Cooper GF, Hennings-Yeomans P, Visweswaran S, *et al.* An efficient bayesian method for predicting clinical outcomes from genome-wide data. *AMIA Annu Symp Proc* 2010;**2010**:127–31.

22 The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;**489**:519–25. doi:10.1038/nature11404

23 Daly R, Shen Q, Aitken S. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review* 2011;**26**:99–157. doi:10.1017/S0269888910000251

24 Gopalakrishnan V, Lustgarten JL, Visweswaran S, *et al.* Bayesian rule learning for biomedical data mining. *Bioinformatics* 2010;**26**:668–75. doi:10.1093/bioinformatics/btq005

25 Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers. *Mach Learn* 1997;**29**:131–63. doi:10.1023/A:1007465528199

26 Kontkanen P, Myllymäki P, Silander T, *et al. On supervised selection of Bayesian networks*. Morgan Kaufmann Publishers Inc. 1999.

27 Jiang X, Cai B, Xue D, *et al.* A comparative analysis of methods for predicting clinical outcomes using high-dimensional genomic datasets. *J Am Med Inform Assoc* 2014;**21**:e312–9. doi:10.1136/amiajnl-2013-002358

28 Yang W, Diamond AM. Selenium-binding protein 1 as a tumor suppressor and a prognostic indicator of clinical outcome. *Biomark Res* 2013;**1**:15–5. doi:10.1186/2050-7771-1-15

Analysis of Viral Genetics for Estimating Diffusion of Influenza A H6N1

Matthew Scotch, PhD, MPH[1], Marc A. Suchard, MD, PhD[2], Peter M. Rabinowitz, MD[3]
[1]Arizona State University, Tempe, AZ, USA; [2]University of California, Los Angeles, Los Angeles, CA; [3]University of Washington, Seattle, WA

Abstract

H6N1 influenza A is an avian virus but in 2013 infected a human in Taiwan. We studied the phylogeography of avian origin H6N1 viruses in the Influenza Research Database and the Global Initiative on Sharing Avian Influenza Data EpiFlu Database in order to characterize their recent evolutionary spread. Our results suggest that the H6N1 virus that infected a human in Taiwan is derived from a diversity of avian strains of H6N1 that have circulated for at least seven years in this region. Understanding how geography impacts the evolution of avian influenza could allow disease control efforts to focus on areas that pose the greatest risk to humans. The serious human infection with a known avian influenza virus underscores the zoonotic potential of diverse avian strains of influenza, and the need for comprehensive influenza surveillance in animals and the value of public sequence databases including GISAID and the IRD.

Introduction

In June, 2013 the Taiwan CDC identified a case of H6N1 influenza A in a 20 year-old female[1]. The index case developed pneumonia, was hospitalized, yet survived. Initial phylogenetic analysis of the viral genome determined that this isolate evolved from chickens in Taiwan[2]. As of September 2014, there has been no documentation of person-to-person transmission of the virus. However, there is still a limited understanding of its phylogeography that might identify vital geographic routes of its genetic lineage. This could enable health agencies to curb future outbreaks of the avian virus and reduce the potential for human-to-human transmission.

Methodology

In order to study the spread of H6N1 avian influenza viruses, we downloaded the entire genome (eight gene segments) of H6N1 avian influenza from the Influenza Research Database (IRD)[3]. We also downloaded sequences of the human H6N1 isolate from the Global Initiative on Sharing Avian Influenza Data (GISAID) EpiFlu database[4]. We created separate FASTA files for each gene and used the strain name to extract geographic and temporal metadata for each stored sequence. For locations outside of China and Taiwan, we altered the definition line of each sequence (indicated by ">") to include the continent rather than the province. We used BEAST v 1.8[5] to perform a Bayesian discrete phylogeography reconstruction[6] of the evolutionary spread of the virus between geographic locations. We then created maximum clade credibility (MCC) trees for each gene from our posterior samples in order to construct single "best" gene trees.

For each gene, we specified a Markov chain Monte Carlo (MCMC) chain length of 30,000,000, sub-sampling every 1,000 steps. We analyzed the effective sample size (ESS) of the parameters using Tracer[7] and, if necessary, re-initiated new chains that we combined with LogCombiner[5]. We used TreeAnnotator[5] to specify an MCC with a 10% burn-in to disregard the initial steps in the MCMC. We used FigTree v. 1.4.2[8] to time-scale the MCC by years and color-code the branches by their most probable geographic state. In addition, we calculated the association index (AI) and parsimony scores (PS) using a program called BaTS to determine if the diffusion of H6N1 is geographically structured[9]. These two statistics test the hypothesis that tips in the tree are no more likely to share the same location (trait) with adjoining taxa than by chance alone[9].

Results

We included the following number of H6N1 sequences in the analysis: 223 PB2 sequences, 221 PB1, 227 PA, 303 HA, 213 NP, 316 NA, 258 M, and 349 NS sequences. In the figure, we show the phylogeographic MCC tree for each influenza gene segment. The time-scaled trees have an x-axis that indicates the years of evolution for the H6N1 virus. The posterior mean estimate of the origin of most of the genes is sometime between 1935-1943. HA (posterior mean: 1925) and PB2 (posterior mean: 1913) are a little earlier than that while NS is a hundred years earlier (posterior mean: 1841). These differences in time could be an indication of reassortment events among the gene segments[6]. We draw an arrow to indicate the human virus. For all genes, the human virus is located within the diversity of avian sequences collected in Taiwan from 1997 - 2010. Seven of the eight genes depict that early in its evolution, H6N1 was most likely to be spreading in North America. For example, other than PB1, at least one of the

two direct ancestors to the origin (*i.e*. root of the tree) are most likely from North America. In most of these genes, the virus shows evidence of local North American spread that ultimately formed a clade. Here, we define a clade as viruses that are grouped together in the tree that share the same ancestor[10]. This is distinct from the diversity of avian sequences collected in Taiwan that includes the human virus. The emergence of this diversity was a result of virus spread across different geographical areas including Europe, China, and Hong Kong. Also, we note the most recent geographic location before the diversification of Taiwanese sequences. For seven of the eight genes, the ancestor was most likely split between Hong Kong (PB2, PA, HA, NP) and Europe (PB1, NA, M). One gene, NS, was most likely descended from Korea.

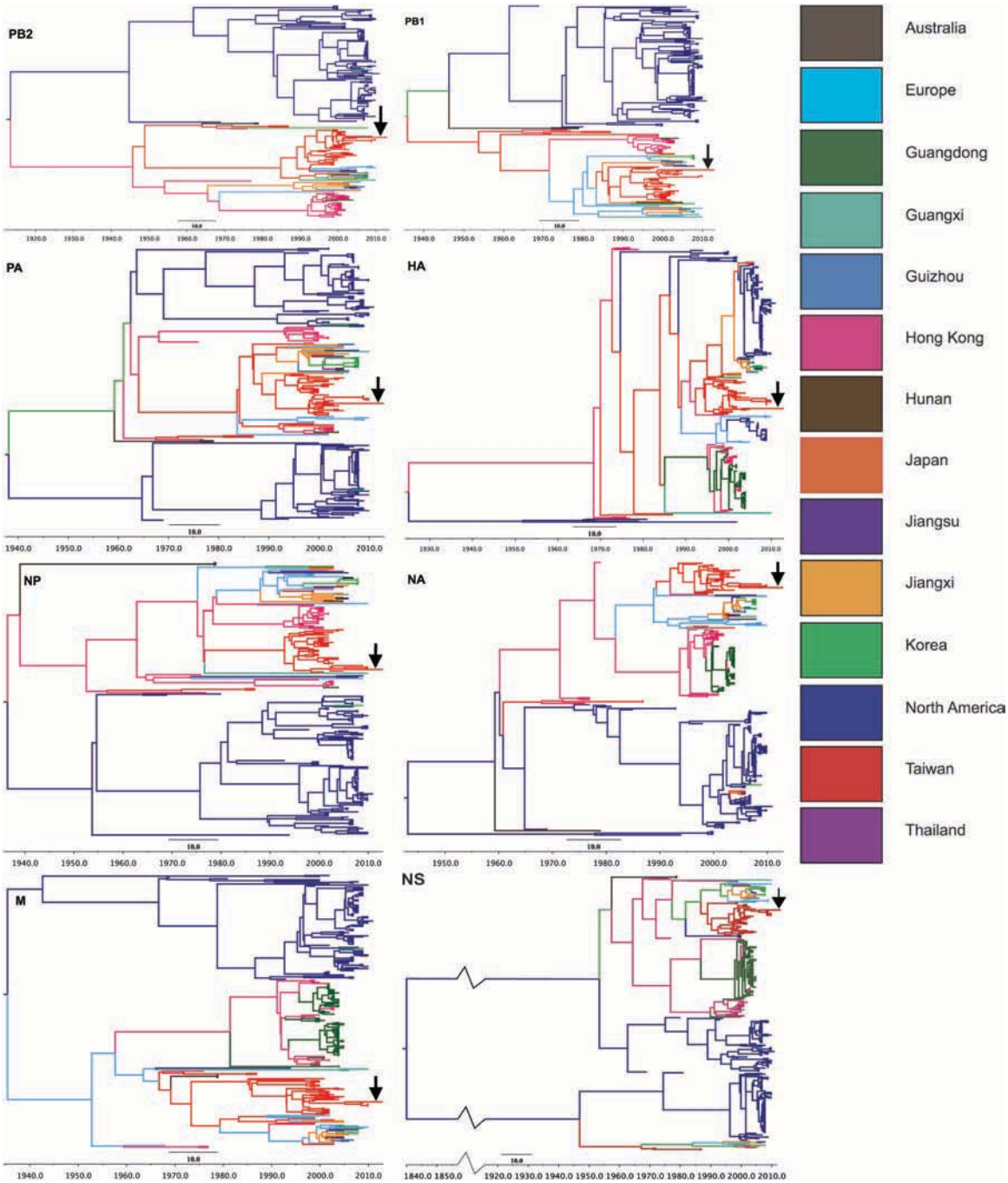


Figure 1. Phylogeographic MCC trees of eight H6N1 influenza A gene segments. Human virus indicated by arrow. Note, we have truncated the long reassortment of the NS segment before the root of the other segments for visualization purposes.

In Table 1, we report the results of the correlation between virus phylogeny and geography by calculating the AI and PS. Here, we present the estimator mean values and the 95% confidence intervals under the observed and null distribution of geographic states. For all genes, the observed values were statistically different from their null estimates, suggesting that the evolution of H6N1 is geographically structured for all gene segments.

In Table 2, we focus solely on the one branch of the tree that contains the human virus. Here, we report the most recent ancestor and year for the human H6N1 virus isolate across all eight gene segments. We determined the most recent ancestor by examining the leaf node of the human virus in the phylogeographic tree (Figure 1). We considered the location with the highest probability to be the most recent ancestor and calculated the Kullback-Leilber score (KL)[6] using the R statistical package[11] to determine the statistical support for the difference between the posterior and prior probability estimates of location. We considered the age of the leaf node to represent the posterior mean year of divergence. We found agreement across all genes for a Taiwan H6N1 strain to be the most recent ancestor with very strong statistical support as indicated by the large KL scores. The age range of the most recent divergence was 1999-2006 suggesting that the virus that infected the individual has been persisting in Taiwan for at least seven years.

Table 1. Association index and parsimony scores for all eight gene segments when considering location as a trait*#

Gene	Association index		Parsimony score	
	Observed Mean	Null Mean	Observed Mean	Null Mean
PB2	1.63 (1.47, 1.86)	16.13 (14.70,17.70)	21.19 (21.00, 22.00)	94.38 (90.43, 97.95)
PB1	2.22 (1.92,2.51)	14.96 (13.30, 16.50)	23.98 (20.00, 25.00)	89.52 (85.98, 92.50)
PA	2.39 (2.06, 2.71)	15.43 (14.00, 16.92	24.67 (24.00, 25.00)	92.39 (89.74, 95.31)
HA	1.48 (1.15, 1.82)	23.90 (22.50, 25.54)	28.57 (27.00, 29.00)	147.00 (143.14, 151.86)
NP	3.09 (2.85, 3.33)	15.73 (14.22, 17.23)	27.92 (27.00, 28.00)	94.39 (90.81, 97.14)
NA	2.30 (1.91, 2.71)	24.03 (22.47, 25.57)	30.52 (29.00, 32.00)	157.05 (152.64, 161.07)
M	2.22 (1.79, 2.66)	21.80 (20.37, 23.10)	29.43 (28.00, 31.00)	138.27 (132.23, 143.55)
NS	1.96 (1.72, 2.27)	27.31 (25.79, 28.66)	29.50 (29.00, 30.00)	172.51 (166.93, 177.27)

*PB2, polymerase basic 2; PB1, polymerase basic 1; PA, polymerase acidic; HA, hemagglutinin; NP, nucleoprotein; NA, neuraminidase; M, matrix; NS, non-structural.

#The p-values for all observed versus mean values are < 0.05 confirming that the diffusion of H6N1 is geographically structured.

Table 2. Location and year of the most recent ancestor of human H6N1 virus for all eight gene segments.

Gene	Most Recent Ancestor (Highest posterior probability)	Posterior Mean Year	Kullback-Leibler divergence
PB2	Taiwan (0.99)	2002	> 7
PB1	Taiwan (0.99)	1999	> 6
PA	Taiwan (0.99)	1999	> 7
HA	Taiwan (0.99)	2003	> 6
NP	Taiwan (0.99)	2005	> 6
NA	Taiwan (0.99)	2005	> 7
M	Taiwan (0.99)	2005	> 6
NS	Taiwan (0.99)	2006	> 8

Conclusion

Our results suggest that the H6N1 virus that infected a human in Taiwan is derived from a diversity of avian strains of H6N1 that have circulated for at least seven years in this region. The vast majority of viruses included in this

diversity were from chickens. In addition, in the HA and NA trees, this diversity included all of the viruses from a Taiwan chicken clade identified by Lee *et al.*[5, 12]. Our finding of Taiwan as the geographic source is consistent with other works on the origin of this virus including Shi *et al.*[13] In addition, we found that North America contributed to the early diffusion of the virus, likely among migratory North American birds, however this has resulted in the formation of a distinct localized clade. Conversely, the formation of the diversity of H6N1 in Taiwan is a result of geographic mixing in Europe and Asian with Hong Kong serving as an important geographic location in the diffusion process.

Understanding how geography impacts the evolution of avian influenza could allow disease control efforts to focus on areas that pose the greatest risk to humans. Epidemiologists can then study local factors including poultry trade and avian migration in order to identify key points for interventions. In addition, the recent cases of human infection in other avian influenza viruses such as H7N9[14] underscores the zoonotic potential of diverse avian strains of influenza, and the need for comprehensive influenza surveillance in animals and the value of public sequence databases including GISAID and the IRD.

Acknowledgements

The project described was supported by award number R00LM009825 from the National Library of Medicine to Matthew Scotch and award number DMS1264153 from the National Science Foundation to Marc A. Suchard. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine, the National Institutes of Health or the National Science Foundation.

We acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu™ Database on which this research is based. The list is detailed below.

Segment ID	Segment	Country	Collection date	Isolate name	Originating Lab	Submitting Lab	Authors
EPI459852	PB2	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459853	PB1	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459854	PA	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459855	HA	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459856	NP	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu;

					Disease Control		Feng-Yee, Chang
EPI459857	NA	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459858	M	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang
EPI459859	NS	Taiwan	2013-May-07	A/Taiwan/2/2013	National Influenza Center, Centers for Disease Control	Taiwan CDC	Ji-Rong, Yang; Ming-Tsan, Liu; Ho-Sheng, Wu; Feng-Yee, Chang

References

1. CDC Taiwan. Laboratory-confirmed case of human infection with avian influenza A(H6N1) virus in Taiwan recovered. 2013 [cited; Available from: http://www.cdc.gov.tw/english/info.aspx?treeid=bc2d4e89b154059b&nowtreeid=ee0a2987cfba3222&tid=E36A5E9AB3D3A216
2. Wei SH, Yang JR, Wu HS, et al. Human infection with avian influenza A H6N1 virus: an epidemiological analysis. The Lancet Respiratory medicine. 2013 Dec;**1**(10):771-8.
3. Squires RB, Noronha J, Hunt V, et al. Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza Other Respi Viruses. 2012 Nov;**6**(6):404-16.
4. GISAID. EpiFlu. 2013 [cited 2013 Apr 6]; Available from: http://platform.gisaid.org/
5. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012 Aug;**29**(8):1969-73.
6. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS computational biology. 2009 Sep;**5**(9):e1000520.
7. Rambaut A, Suchard MA, Drummond AJ. Tracer. 2013 [cited 2014 May 13]; Available from: http://tree.bio.ed.ac.uk/software/tracer/
8. Rambaut A. FigTree. 2013 [cited 2014 May 13]; Available from: http://tree.bio.ed.ac.uk/software/figtree/
9. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases. 2008 May;**8**(3):239-46.
10. Higgs PG, Attwood T. Bioinformatics and Molecular Evolution. Hoboken, NJ, USA: Wiley-Blackwell; 2009.
11. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2014.
12. Lee MS, Chang PC, Shien JH, Cheng MC, Chen CL, Shieh HK. Genetic and pathogenic characterization of H6N1 avian influenza viruses isolated in Taiwan between 1972 and 2005. Avian diseases. 2006 Dec;**50**(4):561-71.
13. Shi W, Shi Y, Wu Y, Liu D, Gao GF. Origin and molecular characterization of the human-infecting H6N1 influenza virus in Taiwan. Protein & cell. 2013 Nov;**4**(11):846-53.
14. Hu J, Zhu Y, Zhao B, et al. Limited human-to-human transmission of avian influenza A(H7N9) virus, Shanghai, China, March to April 2013. Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin. 2014;**19**(25).

Detecting Cancer Pathway Crosstalk with Distance Correlation

Michael F. Sharpnack, Kun Huang, PhD
The Ohio State University, Columbus, OH

Abstract

Biological pathway regulation is complex, yet it underlies the functional coordination in a cell. Cancer is a disease that is characterized by unregulated growth, driven by underlying pathway deregulation. This pathway deregulation is both within pathways and between pathways. Here, we propose a method to detect inter-pathway coordination using distance correlation. Utilizing data generated from microarray experiments, we separate the genes into pathways and calculate the pairwise distance correlation between them. The result is intuitively viewed as a network of differentially dependent pathways. We find intuitive, yet surprising significant hub pathways, including glycophosphatidylinositol anchor synthesis in lung cancer.

Background

Biological pathways in the human cell function together in a highly orchestrated fashion. This coordination results from several mechanisms, including the common occurrence of two metabolic pathways sharing a common substrate. Timing is crucial to the development of the cell and deregulation of the interactions among pathways can have disastrous consequences, such as tumorigenesis. We present a method to detect interactions among pathways from gene expression data of multiple samples, and apply it to identify changes of interactions between pathways in cancers versus normal tissues. We hypothesize that phenotypic changes between two conditions, such as tumor and normal, are associated with changes in pathway dependencies, and further that hub pathways are of special importance to these phenotypic changes. This hypothesis has an advantage that it focuses on the collective behavior of genes in pathways instead of individual genes and therefore does not require correlation between expression profiles of individual genes.

In order to mathematically characterize the functional relationship between two gene lists given their expression profiles over a collection of samples, we implement a relatively new similarity metric called distance correlation[1]. Distance correlation is a type of correlation metric which can detect nonlinear relationships between two vectors or matrices. Given two matrices with the same number of columns, for each matrix, we can consider their columns to be feature vectors for a set of samples. Therefore the distance correlation first calculates the distances between the samples. Then the Pearson correlation coefficient (after a normalization process) between the two sets of distances is computed as the relationship measurement between the two matrices. Geometrically, this is equivalent to compare two weighted networks for the samples and thus exactly matches the notion of our hypothesis.

To test our hypothesis, we develop a two stage workflow. The first stage is to establish a pathway network for samples in different conditions such as cancer versus normal tissues using whole genome transcriptome data from microarray experiments. The second stage is to identify interacting pathways and pathway clusters in specific conditions such as cancers. Our results in multiple cancer studies show that we are able to identify specific pathway interaction in cancers, which supports the notion on altered metabolism processes in cancers. These results suggest that our approach will lead to wide applications as a translational bioinformatics tool for studying diseases at the pathway levels.

Pathway regulation is complex and multifactorial. As such, pathways exhibit both linear and nonlinear dependence on each other. Further complicating the situation, different genes in a pathway have varying levels of importance to the overall function of that pathway. It is not clear what constitutes an active pathway. Some methods have used the average gene expression or a threshold for the number of genes needed to be active to say that the entire pathway is active. One pitfall of these assumptions is in pathways with a highly influential rate-limiting reaction that is controlled by a single enzyme, such as in cholesterol synthesis. Cholesterol synthesis begins with Acetyl-CoA and ends at Cholesterol after six reactions; however, the rate-limiting reaction, the reaction that controls the kinetics of cholesterol synthesis, is the conversion of HMG-CoA to Mevalonate by the enzyme HMG-CoA reductase. This reaction is inhibited by HMG-CoA reductase inhibitors.

Much attention has been given to deregulation of genes within pathways, as this intra-pathway deregulation is the hallmark of many cancers. The question of inter-pathway regulation has only recently been posed, perhaps due to the relatively short time that high throughput expression analysis has been available. The effects of one pathway on another could potentially be as important as the effect of one gene in a pathway on another gene in the same pathway. Many genes exert pleiotropic effects on distinct pathways, and transcriptional regulation can be location-specific, rather than function-specific. In other words, single transcription factors can regulate the genes that belong to distinct functional pathways. Because of this, selecting a single pathway to study is likely an incomplete picture of the causes of tumorigenesis. Pathway coordination, or ``crosstalk", has been studied by calculating differential expression of genes or sets of genes within pathways[2,3,4,5]. These methods also frequently incorporate, and therefore rely on protein-protein interaction networks. The nonlinear nature of our pathway also differentiates it from PAthway Network Analysis (PANA), a method proposed by Ponzoni et al., which can only detect linear patterns[4]. PANA also employs dimensionality reduction methods, which result in a loss of information that our method does not suffer. Cho, et al. use a set-wise interaction score that employs the Renyi relative entropy measure to measure pathway crosstalk; however, methods employing information theory techniques are unlikely to be intuitive to biologists[6]. All of these methods address the question of differentially expressed pathways, not differentially correlated pathways. This subtle difference separates two biological questions. We seek to answer the question, how does the dependency between pathways differ between cancer and normal cells? It is a much more general question than asking whether or not the pathways are over or underexpressed together.

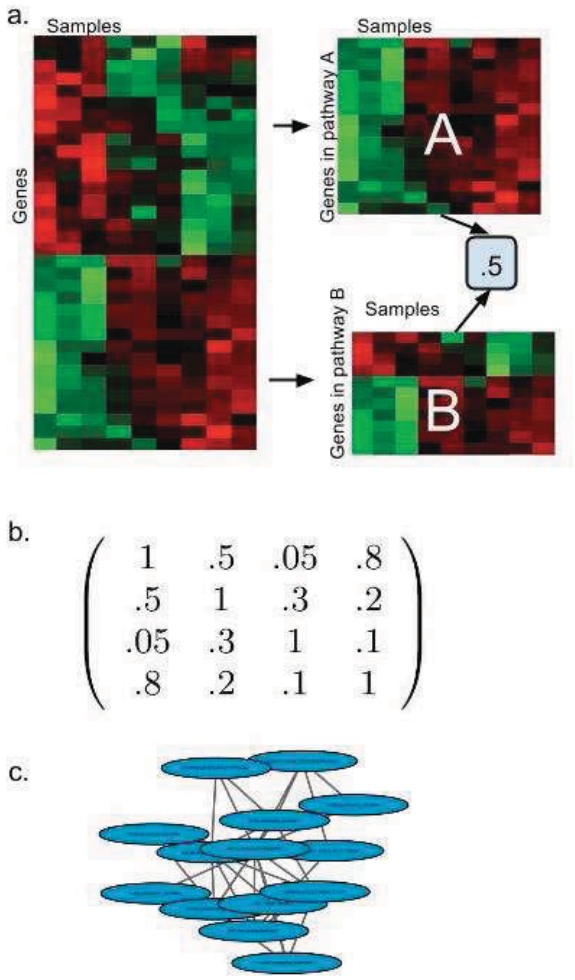


Figure 1 The workflow for establishing pathway dependency networks and compare between different conditions. **(a)** Matrix of expression values for each pathway (A and B) are extracted from the full expression matrix. (**b**) A single distance correlation values represents the dependency on two pathways, such as A and B. These distance correlation values are combined into the symmetric distance correlation matrix shown in **(b)**. **(c)** Two matrices such as that shown in **(b)**, created from two different phenotypes, can be subtracted and their entries used as edge weights to create a differential network. Here we show a representative example of such a network.

Methods

We selected a non-small cell lung cancer paired tumor:normal microarray dataset[7] for study. This dataset was normalized following standard Affymetrix RMA normalization and log transformed. To assign genes to pathways, we adopted the Kyoto Encyclopedia of Genes and Genomes (KEGG)[8] pathways to find the intersection of our microarray genes and genes known to be involved in pathways. In total, we identified 186 diverse pathways from KEGG ranging from cancer-related pathways to signaling and metabolic pathways. Many genes are present in more than one of the 186 pathways, which could prevent a bias in pathway dependence. To minimize the bias between pathways with many genes in common, we simply combine the gene lists of pathways with similar functions and high gene overlap. For instance, 11 pathways related to autoimmunity exhibited high gene overlap, so they were

condensed into a single pathway which is the union of all genes contained in each of 11 pathways. These condensed pathways were relabeled to reflect the overall functional theme. Once the 186 original pathways were condensed, we were left with 116 pathways with low gene overlap. We were able to reduce the number of pathway pairs with >20% genes present in both pathways from 220 to 30. While the remaining 30 pathway pairs had high gene overlap, they were not clearly functionally related, which may be a result of the pleiotropic nature of many proteins. As a further safegaurd against the bias created by pathway overlap, our method considers the differences in dependence, which may be minimal in pathway pairs in which the dependence is high under any experimental conditions due to gene overlap.

To compare the expression of genes within pathways, we employ distance correlation, a measure that can summarize this interaction into a single value. Distance correlation, R, is a measure of the dependence between two random variables, $(X,Y) = \{(X_k,Y_k) : k = 1, \ldots ,n\}$. Conveniently, $0 \leq R \leq 1$, and $R = 0$ if and only if X and Y are independent. In this paper, we consider X_k in $\mathbb{R}^{p,k}$ and Y_k in $\mathbb{R}^{q,k}$ to be microarray expression values for patient k where p and q are the number of genes measured for a given pathways X and Y, respectively. Our expression measurements are therefore organized into 116 p_i x k matrices, where $i = \{1, \ldots 116\}$ and p_i is the number of genes in each pathway. A single distance correlation value is calculated for each pair of pathways, which creates a 116x116 matrix with values between 0 and 1. For our purposes, distance correlation has several advantages when compared to Pearson Correlation: it can detect nonlinear relationships; it can be used to compare two matrices with the same patient sample size but different gene sample size; and it is between 0 and 1. In addition, when a dataset exhibits a bivariate normal distribution, R is a linear function of Pearson Correlation, ρ, with a slope of approximately .9 and $R = |\rho|$ when $|\rho| = 1$. Distance correlation is empirically calculated as described in Szekely et. al[1].

After calculating distance correlation, we arrive at a 116x116 weighted adjacency matrix, where each entry in this matrix is the pairwise distance correlation between two pathways. We have two such matrices, one for tumor samples and one for normal samples. To compare tumor and normal samples, we subtract the normal adjacency matrix from the tumor. We now have the pieces we need to construct a graph representing the change in dependency between each pair of pathways.

To visualize these distance correlation networks, we employed the Clustergram function in Matlab and the widely used network visualization tool, Cytoscape[9]. The Matlab clustergram function uses average linkage for the hierarchical clustering. We then calculated the difference of distance correlation values for every pathway pair between cancer and normal samples for each dataset. The networks were imported into Cytoscape as tables, containing the top .5% of pairwise differential dependencies (based on absolute values of the differences in distance correlation), roughly corresponding to an absolute value of $\geq .15$.

Results

Our method found that the coordination between GPI-anchor biosynthesis and several other pathways, including metabolic pathways, was significantly lower in lung tumor samples than in normal samples as shown **Figure 3** with the clustering of the pathways based on the difference of distance correlation values in **Figure 3.a** and network diagram in **Figure 3.b**. The GPI anchor synthesis, thyroid cancer and circadian rhythm pathways are the three leftmost columns and bottommost rows of **Figure 3.a**, while they are highlighted as yellow nodes in **Figure 3.b**.

Cancer cells are characterized by changes to surface marker proteins, such as glycophosphatidlinositol (GPI)-anchored membrane-bound proteins. For example, carcinoembryonic antigen, a GPI-anchored protein that is usually only expressed in the developing fetus, has been used as a biomarker for colorectal adenoma progression and recurrence[10]. In addition, GPI biosynthetic enzymes have been shown to be elevated in cancer cells [11,12]. We also found that the thyroid cancer pathway was relatively out-of-sync in the cancer samples. Lung and thyroid cancers are diverse, although they may share common pathways. For instance, thyroid transcription factor 1 is active in both lung and thyroid cancers, and its detection is a principal way in which lung adenocarcinomas and large cell carcinomas are differentiated from other lung cancers[13]. We also found that the circadian rhythm pathway was deregulated with several metabolic pathways. Cancer patients are known to have altered circadian rhythms, which is important when considering the timely administration of chemotherapy[14].

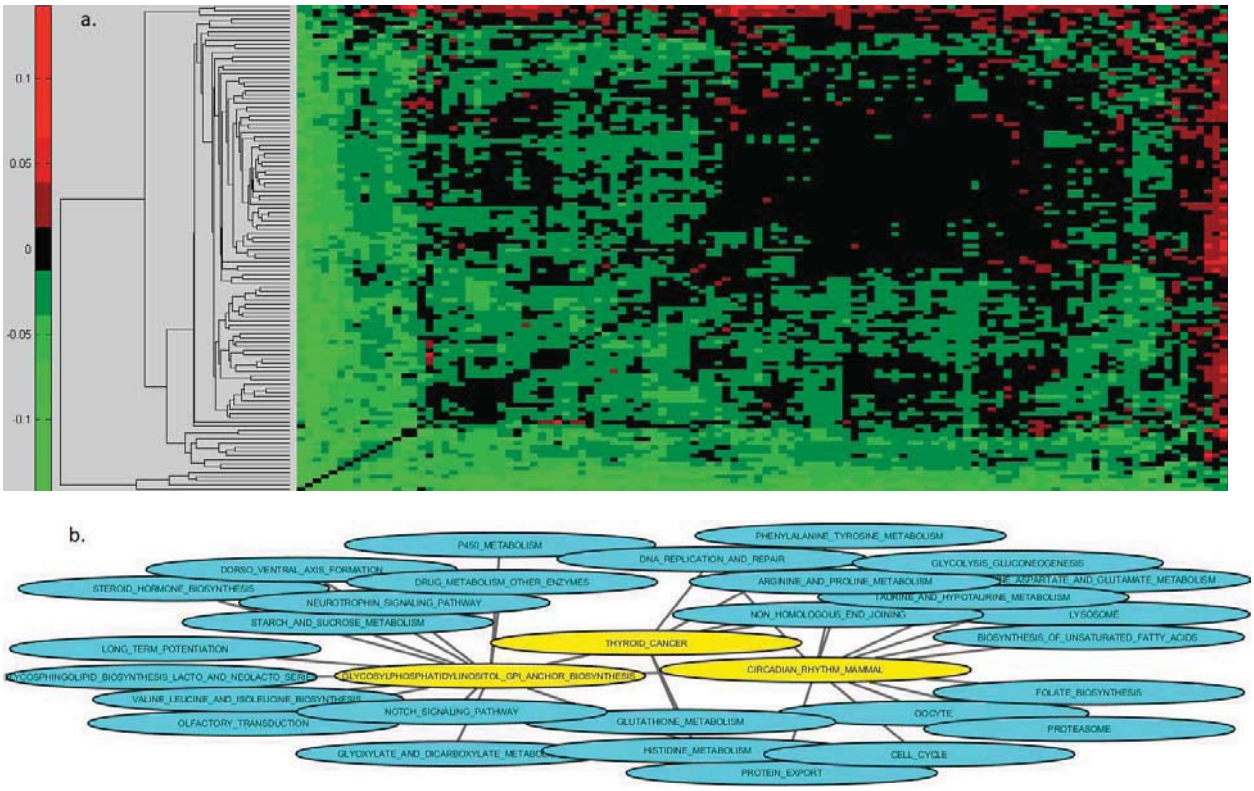


Figure 1. (a) Clustergram of the Lung Cancer pairwise pathway distance correlation with glycophosphatidyl (GPI) anchor synthesis, circadian rhythm, and thyroid cancer pathways in the three leftmost columns and bottommost rows. **(b)** The network diagram of the pathway pairs with high differential dependence, with GPI anchor synthesis, circadian rhythm, thyroid cancer (in yellow) as the clear hub nodes of this network

Discussion

The behavior and fate of a cell can only be understood in the context as an interlocking machine, not as a set of disconnected parts. Pathways are in many ways artificial groupings set up to help humans organize the functions of genes. Enzymes of distinct pathways share cofactors, and the product of one pathway may be the substrate of another. Understanding how these pathways interact is key to identifying the effect on the cell that is created by altering a gene. In this paper, we focus at the interactions among pathways, which provides complementary insights with reduced number of elements in the system. Our method is based on widely available gene expression data. Our use of distance correlations enables multivariate analysis without the need to identify correlated genes. Combining distance correlation and networks also breaks from a long tradition of using differential gene expression to identify important pathways. Our method is simple, requires no parameter input from the user, and it seeks to answer a fundamental question of biology: are two pathways dependent? Using more extensive pathway databases and predesigned datasets, we could explore pathway dependency in greater detail, and perhaps even elucidate the underlying genes responsible for the pathway dependency. The results on a large lung cancer demonstrated the effectiveness of our method in generating new insights on pathway interactions during the disease process.

References

1. Szekely G, Rizzo M, Bakirov N. Measuring and Testing Dependence by Correlation of Distances. The Annals of Statistics 35;6: 2769-94 2007
2. Wang T., Gu J., Yuan J., Tao R., Li Y., Li S. Inferring Pathway Crosstalk Networks Using Gene Set Co-Expression. Molecular BioSystems 9;7: 1822-8 2013
3. Huang Y., Li S. Detection of characteristic sub pathway network for angiogenesis based on the comprehensive pathway network. BMC Bioinformatics 11 Suppl. 1:S32, Jan. 2010

4. Ponzoni I, Nueda M, Tarazona S, Gotz S, Montaner D, Dussaut J, Dopazo J, Conesa A. Pathway Network Inference From Gene Expression Data. BMC Systems Biology 8(Suppl 2):S7, 2014
5. Dutta B, Wallqvist A, Reifman J. PathNet: a Tool for Pathway Analysis Using Topological Information. Source Code for biology and Medicine, 7(1):10, Jan. 2012
6. Cho SB, Kim J, Kim JH. Identifying Set-Wise Differential Co-expression in Gene Expression Microarray Data. BMC bioinformatics 10:109, Jan. 2009.
7. Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R, Farez-Vidal ME. Gene Expression Profiling Reveals Novel Biomarkers in Nonsmall Cell Lung Cancer. International Journal of Cancer 129(2):355--64, July 2011.
8. Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28(1):27-30, Jan. 2000.
9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome research 13(11):2498--504, Nov. 2003.
10. Hammarstrom S. The carcinoembryonic antigen (CEA) family: Structures, Suggested Functions and Expression in Normal and Malignant Tissues. Semin. Cancer Biol., 9(2):67--81.
11. Dolezal S, Hester S, Kirby PS, Nairn A, Pierce M, Abbott KL. Elevated Levels of Glycosylphosphatidylinositol (GPI) Anchored Proteins in Plasma from Human Cancers Detected by C. Septicum Alpha Toxin. Cancer biomarkers 14(1):55--62, Jan. 2014.
12. Wu G, Guo Z, Chatterjee A, Huang X, Rubin E, Wu F, Mambo E, Chang X, Osada M, Sook Kim M, Moon C, Califano JA, Ratovitski EA, Gollin SM, Sukumar S, Sidransky D, Trink B. Overexpression of Glycosylphosphatidylinositol (GPI) Transamidasesubunits Phosphatidylinositol Glycan Class T and/or GPI Anchor Attachment 1 Induces Tumorigenesis and Contributes to Invasion in Human Breast Cancer. Cancer Research, 66(20):9829--36, Oct. 2006.
13. Kaufmann O and Dietel M. Thyroid Transcription Factor-1 is the Superior Immunohistochemical Marker for Pulmonary Adenocarcinomas and Large Cell Carcinomas Compared to Surfactant Proteins A and B. Histopathology 36(1):8--16, 2000.
14. Mormont MC and Levi F. Circadian-System Alterations During Cancer Processes: A Review. International Journal of Cancer 70(2):241--7, Jan. 1997.

Mining Relation Reversals in the Evolution of SNOMED CT Using MapReduce

Shiqiang Tao[1,2], MS, Licong Cui[1], PhD, Wei Zhu[1], Mengmeng Sun[1], Olivier Bodenreider[3], MD, Guo-Qiang Zhang[1,2], PhD
[1]Department of EECS, Case Western Reserve University, Cleveland, OH, USA
[2]Division of Medical Informatics, Case Western Reserve University, Cleveland, OH, USA
[3]National Library of Medicine, Bethesda, MD 20892, USA

Abstract. *Relation reversals in ontological systems refer to such patterns as a path from concept A to concept B in one version becoming a path with the position of A and B switched in another version. We present a scalable approach, using cloud computing, to systematically extract all hierarchical relation reversals among 8 SNOMED CT versions from 2009 to 2014. Taking advantage of our MapReduce algorithms for computing transitive closure and large-scale set operations, 48 reversals were found through 28 pairwise comparison of the 8 versions in 18 minutes using a 30-node local cloud, to completely cover all possible scenarios. Except for one, all such reversals occurred in three sub-hierarchies: Body Structure, Clinical Finding, and Procedure. Two (2) reversal pairs involved an uncoupling of the pair before the is-a coupling is reversed. Twelve (12) reversal pairs involved paths of length-two, and none (0) involved paths beyond length-two. Such reversals not only represent areas of potential need for additional modeling work, but also are important for identifying and handling cycles for comparative visualization of ontological evolution.*

Introduction

The focus of this paper is on ontology evolution [1, 2], most specifically on hierarchical relation reversals in SNOMED CT. A simple example of such a reversal consists of two concepts: "Joint structure of shoulder girdle" and "Joint structure of shoulder region." The 07/2013 version states that "Joint structure of shoulder girdle" is-a "Joint structure of shoulder region," although the (more recent) 03/2014 version asserts the opposite (first two components in Fig. 1): "Joint structure of shoulder region" is-a "Joint structure of shoulder girdle."

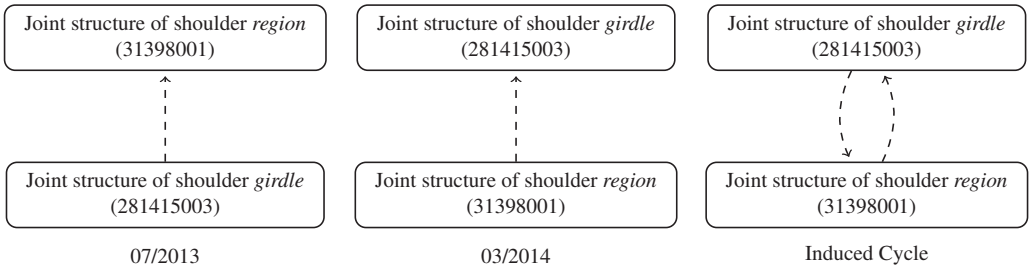


Figure 1: First two arrows: reversal of is-a relation between two versions of SNOMED CT. Right most: a cycle induced by the reversal pair when it appears in a merged graph. The numbers below concept labels are the corresponding SNOMED CT identifiers.

This is an example of a *direct (hierarchical relation) reversal*: "A is-a B" in one version has been changed to "B is-a A" in another version. An *indirect (hierarchical relation) reversal*, considered in this paper, may involve *an arbitrary number of steps* in an entire path: $A \rightarrow^* B$ in one version is changed to $B \rightarrow^* A$ in another version, where $\rightarrow^*$ represents several is-a steps in the same direction. We call the concepts A and B involved in either a direct or an indirect relation reversal *a reversal pair*.

The purpose of this study is twofold: (1) Relation reversals represent an important and rather dramatic structural change, because all the parents and children of the reversed concepts are also affected. There may be good reasons for the occurrence of such reversals that could provide us insight for improving concept labels that better reflect the intended meaning. (2) Relation reversals are important for identifying and handling cycles for comparative visualization of ontological evolution. A common, perhaps most effective, tool for rendering directed acyclic graphs in general, and hierarchical relations in ontological structures in particular, is topological sort (a.k.a. Coffman-Graham algorithm). Topological sort enables each concept assigned a unique *level*, followed by edge rendering. We are interested in visualizing ontological changes in such a way that two related fragments from different versions of the same ontology are *merged* into a single graph for visual inspection of the changes (Fig. 2). However, if a reversal pair is involved (Fig. 1, right most), it causes the merged graph cyclic, making topological sort not directly applicable. By identifying and handling reversals (the only source for introducing cycles) ahead of rendering, topological sort can still be utilized.

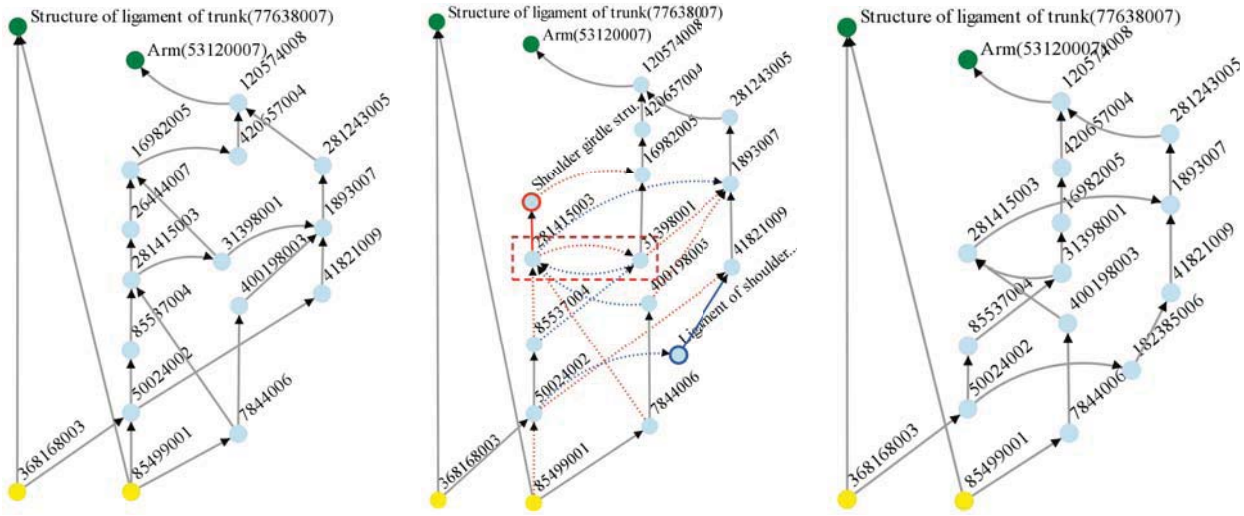


Figure 2: Semi-automatically rendered graphs from two SNOMED CT versions. Left: a non-lattice fragment of 7/2013 version of SNOMED CT. Right: a non-lattice fragment of 03/2014 version of SNOMED CT. Middle: merged graph showing the changes. The loop inside dotted red rectangle is caused by the reversal given in Fig. 1.

Mining all reversals (not just the direct ones) and between all SNOMED CT versions (not just the consecutive versions), is a computationally intensive task. We present a "Big Data" approach using MapReduce to systematically extract all such reversals among 8 SNOMED CT versions from 2009 to 2014. Taking advantage of transitive closure and a MapReduce algorithm to perform large-scale set operations, a total of 48 reversals were found among 8 SNOMED CT versions in 28 pairwise comparisons, using a total of 18 minutes with a 30-node local cloud. The systematic pairwise comparison is necessary to account for all possible situations such as "A is-a B" in a 2009 version has been changed to "B is-a A" in a 2012 version, but A and B are *not related by is-a* for all versions in 2010 and 2011. Except for one, all direct and indirect reversals occurred in three sub-hierarchies: Body Structure, Clinical Finding and Procedure.

1 Background

SNOMED CT. Developed by the International Health Terminology Standard Development Organization (IHTSDO), SNOMED CT is the world's largest clinical terminology and provides broad coverage of clinical medicine, including findings, diseases, and procedures for use in electronic medical records [3]. The international release of SNOMED CT is produced twice a year, reflecting both changes to medical knowledge (e.g., new drugs) and changes to the editorial process (e.g., changes to the representation of anatomical entities). The member countries of the IHTSDO also create extensions of SNOMED CT, with additional concepts specific to the needs of a particular country.

Ontology Evolution. Most ontologies in life science evolve continuously to account for new discoveries, to improve quality, and to align and enrich with related ontologies [4, 5, 6]. Typical ontological changes include the insertion and deletion of concepts as well as the insertion and deletion of relations between concepts. One of the values of evolutionary analysis of ontological structures is to identify regions of more intensive change activities, for targeted ontology quality assurance work [7]. Non-lattice fragments (see Fig. 2) are often indicative of structural anomalies in ontological systems. Such fragments represent possible areas of focus for subsequent quality assurance work [8] because of two reasons: (1) such structures are somewhat incompatible with the generally applicable ontology design principle that the subsumption relationship (is-a hierarchy) should form a lattice [9]; and (2) these fragments have been experimentally validated to represent change rates up to about 38 times higher than those of the overall change [10].

"Big Data" Approach for Ontology Evolution Analysis and Quality Assurance. In our work [10], we introduced MaPLE, an approach using cloud computing to systematically extract non-lattice fragments in 8 SNOMED CT versions from 2009 to 2014, with an average total compute time of less than 3 hours per version. This work used the MapReduce [11] distributed programming environment to process large amounts of data in a scalable way. A MapReduce job consists of a mapper and a reducer function, specified by the user to process data in the form of key-value pairs. Such a job is automatically broken into tasks executed in parallel across a cluster of machines called compute nodes. The results are then aggregated and grouped by the keys by reducer tasks, also executed in parallel. In this

paper we use MapReduce to systematically extract all reversals among 8 SNOMED CT versions from 2009 to 2014, taking advantage of transitive closure and a MapReduce algorithm to perform large-scale set-operations.

2 Methods

Our data source consists of 8 versions of SNOMED CT, dated 07/2009, 01/2010, 01/2011, 01/2012, 07/2012, 01/2013, 07/2013, and 03/2014. To detect direct and indirect hierarchical reversals among these versions of SNOMED CT, we first compute the transitive closure for each version based on the direct "is-a" relationship. Hierarchical relation reversals are then detected by set operations of the respective transitive closures using MapReduce. If all reversal pairs are direct reversal for these 8 versions, we would like to confirm so; but it does not rule out possible indirect reversals occurring between future versions. This is the rationale for using (indirect) transitive closure, to exhaustively detect all direct and indirect reversals, including pairs that are separated by several steps in a path.

Computing Transitive Closure Using MapReduce. On average, each SNOMED CT version contains about 300k concepts and 450k "is-a" relations. Sequential algorithms for computing transitive closure, such as the Floyd-Warshall algorithm, are time-consuming. MapReduce enables a parallel, distributed way to compute transitive closure in a more efficient manner. Fig. 3 (left) is our MapReduce algorithm for computing transitive closure. First, a hash map is setup to load concepts and their direct parents in each computing node using $DistributedCache$. Then, in the map phase (lines 3-16), each mapper reads in a concept, and recursively collects its ancestors level by level, and emits the concept and the set of its ancestors. In the reduce phase (lines 17-21), each reducer emits all concept-ancestor pairs.

MapReduce for Transitive Closure

1: **Input:** Concept nodes and "is-a" relation pairs
2: **Output:** Transitive closure concept pairs

3: **class** MAPPER
4: Setup a HashMap CP and load it with concepts and their direct parents using *DistributedCache*.
5: **method** MAP(concept c)
6: $P = CP$.get(c) ▷ Get direct parents of c
7: $A = \emptyset$ ▷ Initialize a set for ancestors of c
8: **while** $P \neq \emptyset$ **do**
9: A.add(P)
10: $temp = \emptyset$
11: **for** each concept p in P **do**
12: $temp$.add(CP.get(p))
13: **end for**
14: $P = temp$
15: **end while**
16: EMIT(c, A)

17: **class** REDUCER
18: **method** REDUCE(concept c, concept ancestors A)
19: **for** each concept a in A **do**
20: EMIT(c, a) ▷ Output transitive closure pairs
21: **end for**

MapReduce Set Operations for Reversal

1: **Input:** Transitive closure concept pairs for two versions O and N
2: **Output:** Direct and indirect reversals between O and N

3: **class** MAPPER
4: **method** MAP(concept c_1, concept c_2)
5: **if** the concept pair (c_1, c_2) is in O **then**
6: EMIT((c_1, c_2), O)
7: **else if** the concept pair (c_1, c_2) is in N **then**
8: EMIT((c_2, c_1), N) ▷ Reverse the concept pair in N
9: **end if**

10: **class** REDUCER
11: **method** REDUCE((c_1, c_2), versions V)
12: **if** $|V| = 2$ **then** ▷ The concept pair is in both O and reversed N
13: EMIT(c_1, c_2)
14: **end if**

Figure 3: Left - MapReduce steps to compute transitive closure. Right - MapReduce steps to compute reversals.

Performing Big-Set-Operations Using MapReduce. Using the computed transitive closures for each SNOMED CT version, reaching over 5 million edges each, we detect reversals between any two versions by intersecting concept pairs in one version and reversed concept pairs in the other version. Formally, given O and N, transitive closures for two SNOMED CT versions, the set of reversals between them is $\{(c_1, c_2) \mid (c_1, c_2) \in O\} \cap \{(c_2, c_1) \mid (c_1, c_2) \in N\}$. This involves big-set-intersection, since transitive closure for each version contains a large number of concept pairs and traditional way of performing set operations does not always fit into memory. Therefore, we perform big-set-intersections in a more feasible and efficient way using MapReduce. Fig. 3 (right) shows the MapReduce algorithm to perform big-set-intersections and detect reversals between any two SNOMED CT versions. In the map stage (lines 3-9), each mapper reads in a set of concept pairs, and emits key-value pairs $((c_1, c_2), O)$ if the concept pair (c_1, c_2) is in O, and $((c_2, c_1), N)$ if the concept pair (c_1, c_2) is in N. In the reduce stage (lines 10-14), each reducer aggregates versions involved for a concept pair, and emits the concept pair if it belongs to both versions.

3 Results

Relation reversals. We performed 28 pairwise comparisons among the 8 SNOMED CT versions and found 48 reversals (Table 1). Among the 48 reversals, 33 were from the sub-hierarchy Clinical Finding, 8 from Body Structure, 6 from Procedure, and 1 from Event. Two of the reversals (rows 07/2009 → 03/2014 and 01/2010 → 07/2013 in Table 1) had intermediate stages in which the pair is not coupled by an is-a relation, confirming our strategy to perform all

1. Premature or threatened labor (287979001), Premature labor (6383007);
2. Anesthesia for procedure on head and neck (82973008), Anesthesia for procedure on head (120212000);
3. Primary dilated cardiomyopathy (195021004), Primary idiopathic dilated cardiomyopathy (53043001);
4. Computed tomography of shoulder (241564007), Computed tomography arthrogram of shoulder (241583000);
5. Musculoskeletal structure of sacral spine (297169002), Sacral spine (303950008);
6. Rupture of tendon of biceps, long head (86128003), Rupture of tendon of biceps (428883008);
7. Joint structure of shoulder girdle (281415003), Joint structure of shoulder region (31398001);
8. Calcium deposits in tendon (404224009), Osteodesmosis (404225005);
9. Sciatic neuropathy (52585001), Sciatic nerve lesion (367137004);
10. Fly bite (283345006), Mosquito bite (283344005).

Figure 4: 10 sample reversal pairs among the result of 48.

28 pairwise comparisons rather than performing comparison only for 7 consecutive versions.

Ten sample reversal pairs are displayed in Fig. 4. On average, computing the transitive closure of an entire SNOMED CT version and computing big-set-intersection between transitive closures each took less than 40 seconds, amounting to a total computing time of 18 minutes.

Version Pair	C	B	P	E
07/2009 → 01/2010	7	0	2	0
01/2010 → 01/2011	7	0	3	0
01/2011 → 01/2012	7	0	0	0
01/2012 → 07/2012	0	0	0	0
07/2012 → 01/2013	1	0	0	0
01/2013 → 07/2013	1	0	0	0
07/2013 → 03/2014	9	7	1	1
07/2009 → 03/2014	0	1	0	0
01/2010 → 07/2013	1	0	0	0
Total (**48**)	33	8	6	1

Table 1: Distribution of relation reversals in SNOMED CT sub-hierarchies and the versions in which they occurred. C: Clinical Finding. B: Body Structure. P: Procedure. E: Event.

Indirect Reversals. We found 12 indirect reversals. All such pairs involved one direct is-a relation in one version and a length-two path in the other version. Fig. 5 shows 2 such indirect reversals. This confirms the validity of our strategy to compute transitive closures of SNOMED CT versions, because using the direct relations alone would have missed such reversals. Our exhaustive analysis using transitive closure also assured that no reversals involving a path-length of more than 2 existed for the versions we analyzed. However, this does not rule out the existence of indirect reversals involving longer paths between future versions.

Enabling Visualization of Merged Fragments from Distinct SNOMED CT Versions. Our work on detecting direct and indirect reversals is also motivated by removing a technical barrier in visualizing merged ontological fragments from distinct versions of SNOMED CT. Ontological fragments in SNOMED CT can be visualized using SVG (scalable vector graphics - see Fig. 2), supported by common web browsers using the D3 drawing library (http://www.d3js.org). By convention, nodes represent concepts and edges represent "is-a" relation between concepts, with edge direction going from child (lower) to parent (higher). A well-known rendering algorithm for directed acyclic graphs is based on topological sort. We use this algorithm to render merged non-lattice fragments from different versions of SNOMED CT. However, relation reversals introduce cycles, making topological sort non-terminating. Thus detecting reversals before applying topological sort is required.

Fig. 2 illustrates an example of a semi-automatically generated merged graph of two non-lattice fragments from [10] in distinct SNOMED CT versions. The fragments are generated from Body Structure concepts "Arm (53120007)" and "Structure of ligament of trunk (77638007)" in 01/2013 and 03/2014 versions of SNOMED CT. The loop (inside the red dotted rectangle) in the graph involves the reversal pair (item 7 in Fig. 4) of "Joint structure of shoulder girdle" (281415003) and "Joint structure of shoulder region" (31398001).

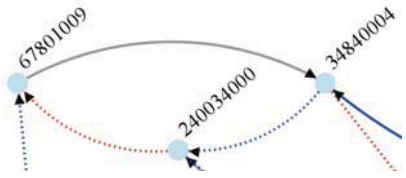


Figure 5: Two pairs of indirect reversals. One pair consists of Tendinitis AND/OR tenosynovitis (240034000) and Inflammatory disorder of tendon (34840004); the other pair consists of Tendinitis AND/OR tenosynovitis (240034000) and Tenosynovitis (67801009).

For Fig. 2, the source concept pair is colored in green. Nodes representing the greatest common descendants of the source pair are painted in yellow. All nodes lying in-between a green node and a yellow node appear in light gray. Additional graphical elements are used for visualizing graph changes (for both Fig. 2 and Fig. 5): red represents deletion and blue insertion. Nodes and edges are also marked with distinct styles to represent additional information: (a) nodes with solid red borders and solid red edges represent deletion: they appear in old fragment but not new; (b) nodes with solid blue borders and solid blue edges represent addition: they appear in new fragment but not old; (c) nodes with dashed borders, and edges drawn in dashed style, represent insertion and deletion in the SNOMED CT versions (such changes must appear in the respective fragments).

Discussion. Clinically, hierarchical relation reversals may involve concepts whose positions in the hierarchical structure are not immediately obvious. Concepts that are the object of relation reversals during ontological evolution may be worth further analysis (such as items 7 to 10 in Table 4). A majority of the reversals are consistent with two prior studies on the use of "and" and "or" [12] as well as "lexically assign, logically refine" [13].

And/Or. We found 18 reversal pairs involved the connectives "and" and "or" (or implicit logical conjunction, e.g., anorectal). For example (see items 1 and 2 in Table 4), "Premature or threatened labor, Premature labor" and "Anesthesia for procedure on head and neck, Anesthesia for procedure on head" are reversal pairs that represent a common possible misinterpretation of "and," which uses intersection in form to represent union in meaning. To make the intended meaning clearer, it is perhaps helpful to normalize the connective to "AND/OR" when the intended meaning is union, especially with respect to Body Structure and Procedure. With this convention, a concept in the form of A should always be a subclass of a concept of the form A AND/OR B. Further, all concepts "A and B" should be normalized to "A AND/OR B," so "Anesthesia for procedure on head and neck" should be normalized to "Anesthesia for procedure on head AND/OR neck" to avoid potential confusion. This is consistent with the analysis given in [12].

Lexically assign, logically refine. Examples due to this type of phenomenon include the pairs in items 3, 4, 5 and 6 in Table 4. The lexicographical difference between the pairs involves the insertion of words, such as "idiopathic" in item 3, "arthrogram" in item 4, "Musculoskeletal structure of" in item 5, and "long head" in item 6. It is arguable that any such insertion of words results in a more specialized concept, and hence should be a subclass of the parent concept. However, in the latest version we analyzed (3/2014), the opposite seems to be the case sometimes. Further consideration is needed to come up with a guiding principle (rule) that can be systematically applied.

Big Data approach for ontology quality assurance work. In this paper we refer to *"Big Data" as a frame of mind, or a "bigger vision,"* in perceiving the scientific landscape from a grander data scale, emboldened by the scalability of cloud computing, such as MapReduce for massive parallel processing. Such an approach can dramatically accelerate the speed of analysis in cases of complex tasks that are less computationally feasible [10, 14]. We believe that such a scalable approach is beneficial for ontology quality assurance work in general, even for computationally feasible problems (such as the work presented here), because it allows us to ask bigger questions and to answer them faster, putting computational barrier on the back of our minds so we can focus more on the scientific content.

Conclusion. We presented a scalable and generalizable method using MapReduce to mine reversals during ontological evolution. 48 hierarchical reversals have been found in 8 SNOMED CT versions from 2009. Identification of such reversals allowed avoidance of cycles in applying topological sort for rendering merged ontological graphs for visual comparison and change illustration. The reversals confirmed prior findings in the literature about concept labeling convention recommendations, but also revealed new cases for further consideration. In general, our closure-based technique has shown to be powerful and efficient for analyzing large ontological structures as well as their evolution. Although this investigation focused on hierarchical reversals, our approach suggests the exploration of reversals of other kinds of relations, which we plan to address in future work.

Acknowledgement. The authors acknowledge partial support by the Case Western Reserve University CTSA Grant UL1TR000439 and in part by the National Library of Medicine throughout its Intramural Research Program.

References

[1] Hartung M, Kirsten T, Gross A, Rahm E. OnEX: Exploring changes in life science ontologies. BMC Bioinformatics. 2009 Aug 13;10:250.

[2] Ceusters W. Applying Evolutionary Terminology Auditing to SNOMED CT. AMIA Annu Symp Proc. 2010 Nov 13;2010:96-100.

[3] Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform Vol. 121, pages 279-90, 2006.

[4] Groß A, Hartung M, Prüfer K, Kelso J, Rahm E. Impact of ontology evolution on functional analyses. Bioinformatics. 2012 Oct 15;28(20):2671-7.

[5] Hartung M, Grob A, Rahm E. COnto-Diff: generation of complex evolution mappings for life science ontologies. J Biomed Inform. 2013 Feb;46(1):15-32.

[6] Kirsten T, Gross A, Hartung M, Rahm E. GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. J Biomed Semantics. 2011 Sep 13;2:6. doi: 10.1186/2041-1480-2-6.

[7] Zhu X, Wei JW, Baorto D, Weng C, Cimino J. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomedical Informatics, Vol. 42, pages 412-25, 2009.

[8] Zhang GQ and Bodenreider O. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. American Medical Informatics Association (AMIA) Annual Symposium, pages 922-926, 2010.

[9] Zweigenbaum P, Bachimont B, Bouaud J, Charlet J, Boisvieux JF. Issues in the structuring and acquisition of an ontology for medical language understanding. Methods Inf Med. Vol. 34(1-2), pages 15-24, 1995.

[10] Zhang GQ, Zhu W, Sun M, Tao S, Bodenreider O, Cui L. MaPLE: A MapReduce Pipeline for Lattice-based Evaluation of SNOMED CT. IEEE International Conference on Big Data, 2014;754-9.

[11] Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. OSDI, 2004.

[12] Mendona EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting "and" and "or" in terminology systems. Proc AMIA Symp. 1998:790-4.

[13] Dolin RH, Huff SM, Rocha RA, Spackman KA, Campbell KE. Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies. J Am Med Inform Assoc. 1998 Mar-Apr;5(2):203-13.

[14] Zhu W, Zhang GQ, Tao S, Sun M, Cui L. NEO: Systematic Non-Lattice Embedding of Ontologies for Comparing the Subsumption Relationship in SNOMED CT and in FMA Using MapReduce. AMIA Joint Summits 2015.

Adverse Drug Events-based Tumor Stratification for Ovarian Cancer Patients Receiving Platinum Therapy

Chen Wang, PhD[1], Michael T. Zimmermann, PhD[1], Christopher G. Chute, MD, Dr. PH[1], Guoqian Jiang, MD, PhD[1],
[1]Department of Health Sciences Research, Mayo Clinic, Rochester, MN

Abstract

The underlying molecular mechanisms of adverse drug events (ADEs) associated with cancer therapy drugs may overlap with their antineoplastic mechanisms. In a previous study, we developed an ADE-based tumor stratification framework (known as ADEStrata) with a case study of breast cancer patients receiving aromatase inhibitors, and demonstrated that the prediction of per-patient ADE propensity simultaneously identifies high-risk patients experiencing poor outcomes. In this study, we aim to evaluate the ADEStrata framework with a different tumor type and chemotherapy class – ovarian cancer treated with platinum chemotherapeutic drugs. We identified a cohort of ovarian cancer patients receiving cisplatin (a standard platinum therapy) from The Cancer Genome Atlas (TCGA) (n=156). We demonstrated that somatic variant prioritization guided by known ADEs associated with cisplatin could be used to stratify patients treated with cisplatin and uncover tumor subtypes with different clinical outcomes.

1 Introduction

Ovarian cancer is one of leading causes of cancer death among women in the United States. About 70% of patients at diagnosis present with advanced-stage and high-grade serous ovarian cancer (1). Platinum-based chemotherapy is a standard treatment following a cytoreductive surgery, however, approximately 25% of patients develop platinum-resistance within six months and almost all patients with recurrent disease ultimately develop platinum resistance (2). In addition, partly due to the lack of successful treatment strategies, the overall five-year survival rate for high-grade serous ovarian cancer is only 31%. Although several mechanisms have been revealed to contribute to chemotherapy response (3-5), there are no valid clinical or molecular markers that effectively predict the chemotherapy response.

Recently, the cancer research community is actively working on compiling cancer genomic information, and investigating new therapeutic options and tailored treatment for individual patient according to personal tumor genome. A notable example is The Cancer Genome Atlas (TCGA) research network (6, 7). TCGA has released an ovarian cancer dataset containing a large (for genomics) sample size, comprehensive genomic profiles and clinical outcome information (1). The dataset has been utilized to analyze chemotherapeutic response in ovarian cancers in several previous studies (8, 9).

Adverse drug events (ADEs) are a critical factor for selecting cancer therapy options in clinical practice. For example, cisplatin and carboplatin are two commonly used chemotherapy drugs in the treatment of ovarian cancer and are also used to treat other cancer types. In comparison with cisplatin, the greatest benefit of carboplatin is its reduced side effects, particularly the elimination of nephrotoxic effects (4). These side effects have been well documented in the United States Food and Drug Administration (FDA) Structured Product Labels (SPLs). The underlying molecular mechanisms of adverse drug events (ADEs) associated with cancer therapy drugs may also overlap with their antineoplastic mechanisms. Specifically, that the antineoplastic mechanism of action, which kills tumor cells, may be the same mechanism by which healthy cells are damaged leading to toxicity. In a previous study, we developed an ADE-based tumor stratification framework (known as ADEStrata) with a case study of breast cancer patients receiving aromatase inhibitors (10), and demonstrated that the prediction of per-patient ADE propensity simultaneously identifies high-risk patients experiencing poor outcome.

In the present study, we aim to evaluate the feasibility of the ADEStrata framework with a different tumor type and class of therapy – ovarian cancer treated with platinum chemotherapeutic drugs. We first identified a cohort of ovarian cancer patients receiving cisplatin drugs from TCGA, and retrieved somatic mutations for each patient case. We then conducted variant prioritization that was guided by known ADEs of cisplatin represented by Human Phenotype Ontology (HPO) terms. We performed pathway-enrichment analysis and hierarchical clustering, which identified two patient subgroups. We finally conducted a clinical outcome association study to investigate whether the patient subgroups are significantly associated with survival outcome in univariate and multivariate analysis.

2 Materials and Methods

2.1 Materials

2.1.1 SIDER: A Side Effect Resource

The SIDER (SIDe Effect Resource) is a public, computer-readable side effect resource that contains reported adverse drug reactions (11). The information is extracted from public documents and package inserts; in particular, from FDASPLs. In the present study, we utilized the latest version SIDER 2 that was released on October 17, 2012.

2.1.2 HPO: Human Phenotype Ontology

The HPO project aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human diseases (12). The ontology contains more than 10,000 terms and equivalence mappings to other standard vocabularies such as MedDRA and UMLS. In the present study, we used the latest version of HPO-MedDRA mapping file that is publicly available from the HPO website (13).

2.1.3 eXtasy: A Variant Prioritization Tool

eXtasy is a variant prioritization pipeline developed at the University of Leuven, for computing the likelihood that a given nonsynonymous single nucleotide variants (nSNVs) is related to a given phenotype (14, 15). The eXtasy pipeline takes a Variant Call File (VCF) and one or more gene prioritization files. Each prioritization file is pre-computed for a specific phenotype (HPO term). In the present study, we downloaded and installed the tool on a local Ubuntu server.

2.1.4 TCGA Data Portal

TCGA Data Portal provides a platform for researchers to search, download, and analyze data sets generated by TCGA consortium (16). As of September 2014, there are 586 cases of ovarian serous cystadenocarcinoma (OV) with data. In the present study, we utilized the OV clinical data (including clinical drug data and follow-up data) and somatic mutation data through the Open Access data tier.

2.2 Methods

2.2.1 Identifying HPO ADE Terms Relevant to Platinum Drugs

We first mapped the ADE terms represented in MedDRA UMLS concept unique identifiers (CUIs) from the SIDER 2 database file to the HPO terms using an HPO-MedDRA mapping file produced by HPO development team. Second, we annotated those HPO terms with a flag using the eXtasy HPO term list to indicate whether a HPO-based ADE term can be processed by eXtasy or not. Third, we retrieved those entries (with drug-ADE pairs) using the drug name “cisplatin” and identified a list of ADEs with their HPO term annotations.

2.2.2 Identifying Patient Cohorts by Platinum Drugs and Somatic Mutations from TCGA

We utilized the clinical drug information file of the OV patients from TCGA data portal through its Open-Access HTTP Directory. The spelling corrections were taken for all variants of the three drugs to maximize the sample size of the patient cases. We then identified a set of patient cases (represented by patient barcodes) that were prescribed for the cisplatin.

We also downloaded the somatic mutation file of the OV patients from TCGA data portal in a Mutation Annotation Format (MAF). The format is a tab-delimited file containing somatic mutations for each patient. As eXtasy requires a VCF file as input, we converted the MAF file into a collection of VCF files. Each VCF file contains somatic mutations for a single patient tumor sample. We combined all VCF files for all cisplatin cases into a single VCF file using the patient barcodes identified in the step above.

2.2.3 Variant Prioritization Using HPO ADE Terms

As mentioned above, we installed an instance of the eXtasy tool in a local server and ran the tool with a custom Ruby script. The input consists of a VCF file and a set of pre-computed gene prioritization files for those phenotypes represented by the HPO ADE terms of interest. The output is a file with likelihood scores for input variants of impacting an individual HPO term (17). The scores represent the probability that a variant is high-ranking in all different phenotypes comparing against a null distribution of random rankings. To shed some lights on how the variants could potentially affect protein function, we first classified the input variants into three functional impact categories, calling a variant “high” if it is a frameshift, nonsense, nonstop, or splice-site; and “medium” if it is a missense; and “silent” if it is a mutation not causing protein coding changes. And then we analyzed the function of those variants scored by eXtasy for cisplatin-related HPO terms.

2.2.4 Tumor Mutation Stratification and Clinical Outcome Association Studies

We first selected statistically significant variants based on the eXtasy order statistics (pseudo p-value <0.05). Second, we aggregated genes affected by these prioritized variants across 1,320 canonical pathways collected from the Molecular Signature Database (MSigDB) (18, 19). In order to reduce false discoveries, multiple criteria were applied to further filter out less relevant pathways (binomial distribution p-value >0.05) or pathways containing too few genes (<10 genes). We excluded pathways with less than 10 genes, based on the consideration that small pathways are often subcomponents of larger pathways, and inclusion of them tends to introduce unnecessary

redundancy. Third, we performed hierarchical clustering to highlight pathway-level patterns among cisplatin-treated patients.

We used overall survival (OS) time (years) as a clinical endpoint to measure the outcome of TCGA patients in the identified cohort. We performed both univariate analysis and multivariate cox-regression to assess the association of clusters (produced by hierarchical clustering) with survival. In multivariate analysis, patient age and tumor stage were adjusted for to evaluate the independent outcome-prediction contribution of found tumor cluster. We also analyzed the distribution of patient age and tumor stage in the clusters identified.

3 Results

In total, we identified a list of cisplatin-induced ADEs represented in 95 unique HPO Ids. Of them, 73 HPO Ids (76.8%) are covered in eXtasy tool. Table 1 shows a list of such ADEs relevant to renal toxicity.

Table 1. A list of cisplatin-induced ADEs relevant to renal toxicity represented in HPO terms.

MedDRA UMLS CUI	MedDRA Label	HPO Id	HPO Label	eXtasy
C0341697	Renal impairment	HP:0000082	Abnormality of renal physiology	YES
C0740394	Hyperuricaemia	HP:0002149	Hyperuricemia	YES
C0235416	Blood uric acid increased	HP:0002149	Hyperuricemia	YES
C1565489	Insufficiency renal	HP:0000083	Renal failure	YES
C0035078	Renal failure	HP:0000083	Renal failure	YES
C0020625	Hyponatraemia	HP:0002902	Hyponatremia	YES
C0595916	Nephropathy toxic	HP:0000112	Nephropathy	YES
C0020598	Hypocalcaemia	HP:0002901	Hypocalcemia	YES
C0151723	Hypomagnesaemia	HP:0002917	Hypomagnesemia	YES
C0020621	Hypokalaemia	HP:0002900	Hypokaliemia	YES
C0151747	Renal tubular disorder	HP:0000091	Abnormality of the renal tubule	YES
C1287298	Urine output	HP:0011036	Abnormality of renal excretion	YES
C0032617	Polyuria	HP:0000103	Polyuria	YES

We were able to identify a cohort of 156 OV patients receiving cisplatin treatment from TCGA OV clinical drug data. Of them, 92 OV patients had somatic mutations identified from OV somatic mutation data.
The eXtasy program ignores silent variants. Of the remaining variants, 12% are of high impact (see section 2.2.3) and almost assuredly affect the normal physiologic function of the affected gene. Of the variants scored by eXtasy for cisplatin-related HPO terms, 40% are highly conserved among placental mammals. Because of lack of conservation at many variant sites, approximately 60% cannot be evaluated with common prioritization tools such as SIFT or PolyPhen2. Of those that are evaluable, both SIFT and PolyPhen2 predict 60% of them as deleterious (predictions are 76% concordant). Variants were prioritized for each patient across the ADE phenotypes represented by 73 HPO terms, producing aggregate prioritization scores (max and order statistics).

By hierarchical clustering, 2 distinct patient clusters, organized by pathways (affected by prioritized variants), were identified and are displayed in Figure 1 containing 16 and 76 patients each. Table 2 shows the results of the univariate and multivariate cox-regression analysis for the three clusters. We found that Cluster 2 has a relatively large number of patients (n=76), and is significantly association with poorer survival time in both univariate and multivariate analysis. Table 3 shows the distribution of age and stage in the 2 clusters identified. There is no significant association between the 3 clusters and age/stage, although we noticed that Cluster 2 is enriched with more Stage IIIC and Grade 3 patient cases. Figure 2 shows a Kaplan-Meier plot of survival time for the 2 clusters, derived from our pathway-level analysis, indicating Cluster 2 had the worse survival outcome associated.

4 Discussion

While TCGA catalogs a large number of OV samples, sample size for individual chemotherapies may be small. Thus, we focus first on the most common chemotherapy regimen so that the subgroup of interest is still reasonably large. In our previous study we considered patients receiving aromatase inhibitors (10). Aromatase inhibitors block conversion of precursor hormones to estradiol, effectively turning off the growth signal for estrogen-dependent tumors. Evidence exists for tumor addiction; that loss of this dependent growth signal leads to apoptosis. The healthy tissues most likely to be affected by this treatment are those who routinely use the aromatase enzyme or estrogen signaling in their normal physiology. In this study, we consider a platinum-based therapy whose mechanism of action is to nonspecifically damage DNA. Any cell could be affected. The tissues most affected are those who are quickly growing and have a greater fraction of their DNA accessible. These include the cancer itself,

but also hematologic stem cells and those of the digestive tract. The mechanistic link to the studied ADEs is clearer – kidneys become compromised due to higher blood protein levels and blood cells cannot be replaced as quickly. The therapy's molecular mechanism is responsible for the ADEs considered. The rationale behind nonspecific chemotherapies, such as cisplatin, is to damage tumor cells more than healthy cells, but damage to both is expected.

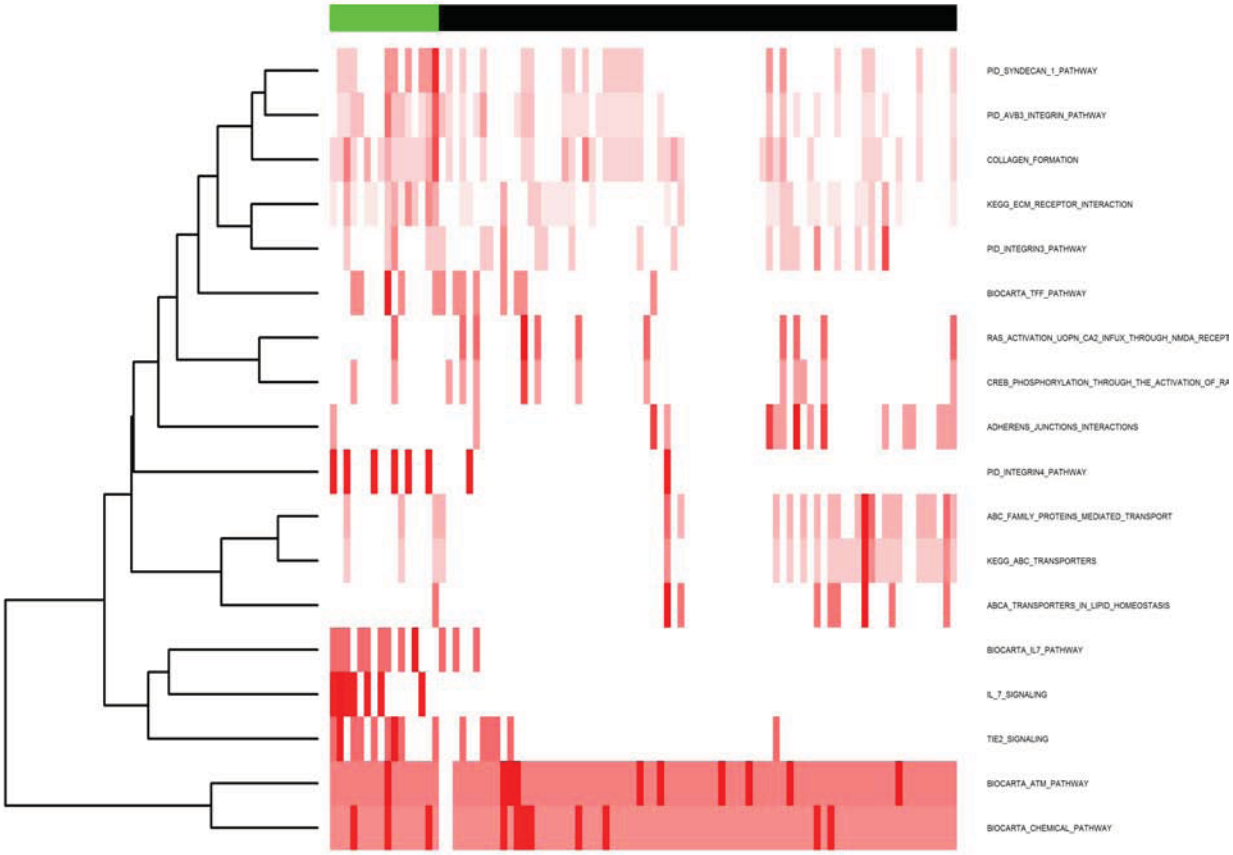


Figure 1. An ordered heatmap showing pathway-level clustering of 92 patients treated with cisplatin across ADE relevant variants. The color of heatmap from white to red indicates low to high percentages (0% to 100%) of genes affected by ADE relevant variants. Column color-bar on top of the heatmap indicates two clusters of samples: Cluster 1 (green) and Cluster 2 (black). Note that the number of the patients (n=92) with pathway enrichment is less than total number of the identified cohort (n=156) is because not all patients have prioritized variants listed.

Table 2. The univariate and multivariate cox-regression analysis results of cluster labels. In multivariate analysis, patient diagnosis age, tumor-grade and tumor-stage were adjusted for to determine the independent contribution of cluster membership. HR denotes hazard ratio; * denotes p<0.05.

Univariate analysis	p-value	HR [95% CI]
Cluster-2	0.019*	3.16 [1.21, 8.30]
Multivariate analysis		
Cluster-2	0.013*	3.47 [1.30, 9.23]
Diagnosis age	0.67	0.99 [0.96, 1.02]
Grade-2	0.18	0.30 [0.05, 1.71]
Grade-3	0.41	0.53 [0.13, 2.34]
Stage-IIIC	0.93	0.96 [0.39, 2.36]
Stage-IV	0.78	0.85 [0.28, 2.61]

Table 3. The distribution of age tumor-grade, and tumor-stage in the two clusters identified. # p-value for age vs. cluster association was computed using ANOVA test; p-value for stage/grade vs. cluster association was computed using Fisher's exact test.

	Cluster-1 (n=16)	Cluster-2 (n=76)	p-value#
Diagnosis age Mean [Q1, median, Q3]	54.3 [47.1, 53.0, 64.4]	56.2 [49.1, 56.2, 61.3]	0.49
Stage (case number)			1
II, IIIA or IIIB	2	8	
IIIC	12	56	
IV	2	12	
Grade (case number)			1
Missing	0	3	
G2	1	8	
G3	15	65	

A logical extension of our current methodology would be to independently predict ADEs given germline or somatic variants. High propensity of ADEs from germline alone would predict high toxicity, while high ADE propensity from somatic variants would point to high efficacy. In a given patient, the ideal

situation would be a prediction of low toxicity and high efficacy, while prediction of high toxicity and low efficacy may be a contra-indication for the therapy. An important implication of our findings in this study is that cisplatin could be more toxic than carboplatin but for a subset of patients it could be more effective. We will pursue retrospective validation of this methodology with the long term goal of aiding clinical decision making in personalized cancer treatment.

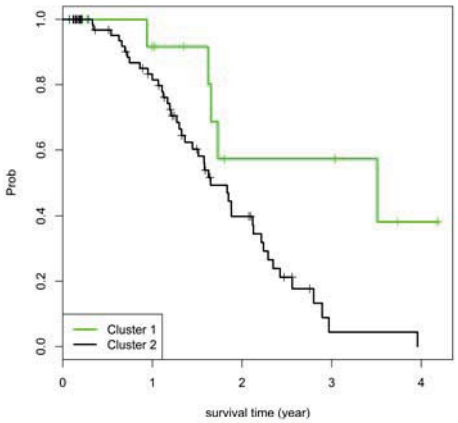


Figure 2. Kaplan-Meier plot of survival time for patients in 2 pathway-level clusters.

5 Conclusion

In summary, we evaluated the feasibility of ADEStrata framework with a different tumor type and chemotherapy class – ovarian cancer treated with platinum chemotherapeutic drugs. We demonstrated that somatic variant prioritization guided by known ADEs associated with cisplatin could be used to stratify patients treated with cisplatin and uncover tumor subtypes with different clinical outcomes. In the future, we plan to evaluate and validate our approach by incorporating more data types (e.g., germline variants), and investigate the generalization of the method in other tumor types.

Reference

1 Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 30;**474**(7353):609-15.
2 Pujade-Lauraine E, Hilpert F, Weber B, et al. Bevacizumab combined with chemotherapy for platinum-resistant recurrent ovarian cancer: The AURELIA open-label randomized phase III trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014 May 1;**32**(13):1302-8.
3 Ow GS, Ivshina AV, Fuentes G, Kuznetsov VA. Identification of two poorly prognosed ovarian carcinoma subtypes associated with CHEK2 germ-line mutation and non-CHEK2 somatic mutation gene signatures. Cell Cycle. 2014 Jul 15;**13**(14):2262-80.
4 Dasari S, Bernard Tchounwou P. Cisplatin in cancer therapy: Molecular mechanisms of action. European journal of pharmacology. 2014 Oct 5;**740C**:364-78.
5 Bowden NA. Nucleotide excision repair: why is it not used to predict response to platinum-based chemotherapy? Cancer letters. 2014 May 1;**346**(2):163-71.
6 The Cancer Genome Atals. [cited February 17, 2014]; Available from: http://cancergenome.nih.gov/
7 Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics. 2013 Oct;**45**(10):1113-20.
8 Bosquet JG, Marchion DC, Chon H, Lancaster JM, Chanock S. Analysis of chemotherapeutic response in ovarian cancers using publicly available high-throughput data. Cancer research. 2014 Jul 15;**74**(14):3902-12.
9 Hsu FH, Serpedin E, Hsiao TH, Bishop AJ, Dougherty ER, Chen Y. Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. BMC genomics. 2012;**13 Suppl 6**:S13.
10 Wang C, Zimmermann MT, Prodduturi N, Chute CG, Jiang G. Adverse Drug Event-based Stratification of Tumor Mutations: A Case Study of Breast Cancer Patients Receiving Aromatase Inhibitors. AMIA Annual Symposium. Wahsington, DC: (in press); 2014.
11 Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. Molecular systems biology. 2010;**6**:343.
12 Kohler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic acids research. 2014 Jan 1;**42**(1):D966-74.
13 The Human Phenotype Ontology URL. [cited February 17, 2014]; Available from: http://www.human-phenotype-ontology.org/
14 Sifrim A, Popovic D, Tranchevent LC, et al. eXtasy: variant prioritization by genomic data fusion. Nature methods. 2013 Nov;**10**(11):1083-4.
15 eXtasy URL. [cited February 15, 2014]; Available from: http://homes.esat.kuleuven.be/~bioiuser/eXtasy/
16 TCGA Data Portal. [cited February 17, 2014]; Available from: https://tcga-data.nci.nih.gov/tcga/
17 Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. Nature biotechnology. 2006 May;**24**(5):537-44.
18 Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;**27**(12):1739-40.
19 Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005 Oct 25;**102**(43):15545-50.

Ranking Medical Subject Headings using a factor graph model

Wei Wei[1], Dina Demner-Fushman[2], Shuang Wang[1], Xiaoqian Jiang[1], Lucila Ohno-Machado[1]
[1]Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA 92093 USA,
Email: {w2wei, shw070, x1jiang, lohnomachado}@ucsd.edu
[2]National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, 20894
Email: ddemner@mail.nih.gov

Abstract

Automatically assigning MeSH (Medical Subject Headings) to articles is an active research topic. Recent work demonstrated the feasibility of improving the existing automated Medical Text Indexer (MTI) system, developed at the National Library of Medicine (NLM). Encouraged by this work, we propose a novel data-driven approach that uses semantic distances in the MeSH ontology for automated MeSH assignment. Specifically, we developed a graphical model to propagate belief through a citation network to provide robust MeSH main heading (MH) recommendation. Our preliminary results indicate that this approach can reach high Mean Average Precision (MAP) in some scenarios.

INTRODUCTION

MeSH is a controlled vocabulary thesaurus used at the NLM for indexing biomedical literature. MeSH indexing improves literature retrieval and it is widely used in biomedical text mining (1). The NLM indexers generally assign 5 to 15 MH to every article using their domain knowledge (2). This process is assisted by the automated MTI system developed at the NLM (3,4). Currently, around 65% of MHs are suggested by MTI.

Automatically assigning MHs was recently a task in the international BioASQ challenge (5). The two winning teams were able to improve the strong baseline provided by MTI using supervised machine learning methods; particularly, learning to re-rank the original MTI results (6). We were therefore motivated to explore a novel machine learning method for finding relevant MHs and providing more accurate suggestions. We investigated a factor graph based approach that uses the hierarchical structure of the MeSH ontology and semantic distance metrics in a case study. The performance of our preliminary model is close to MTI, which considers more attributes such as the abstract and the title.

BACKGROUND

MTI generates a ranked list of MHs, Subheadings and CheckTags as a final result from the title and the abstract of every article, using a combination of two indexing methods: PubMed Related Citations and MetaMap indexing (7). PubMed Related Citations, an implementation of the *k*-Nearest Neighbor (*k*-NN) algorithm, produces a list of related articles; MetaMap maps text from the titles and abstracts of the articles to the UMLS Metathesaurus. The results of the two methods are then clustered and ranked through a post-processing phase (3). After that, the indexers review MTI suggestions and select the appropriate main headings from the pool. The *k*-NN algorithm on its own has outperformed several other approaches in the experiments on 1000 randomly selected MEDLINE citations (8). There are re-ranking algorithms applied to the MTI output (6) to adjust suggestions. Similar ideas have been also applied to the results of the PubMed Related Citations algorithm (9), as well as the multi-label ensemble method consisting of the SVM classifier and the Latent Dirichlet Allocation (LDA) model (10). These algorithms can outperform the strong MTI baseline by about 10%.

Graphic models, which generally include directed graphs, undirected graphs, and factor graphs, are powerful tools for representing probabilistic models. Factor graphs, which can represent both directed and undirected graphs, have gained popularity as they offer great flexibility for problem solving. Although both directed and undirected graphs allow representation of a global function with multiple variables as a product of factors over subsets of variables, a factor graph provides an explicit way to factorize the global function by introducing variable and factor *nodes*. The introduction of factor nodes allows the optimization algorithms (e.g., belief propagation) to be derived in a simple and general form, because the factor graph unifies the directed and undirected graph with the same representation. A factor graph is a bipartite graph that consists of two types of nodes, i.e., factor and variable nodes. Each factor node needs to connect at least one variable node and vice versa. A variable node represents a hidden variable or an observation in an inference problem. A factor node captures the factor function of the variables in the connected variable nodes. For an acyclic graph, a factor function corresponds to a factor of the decomposed joint probability. For example, Figure 1 depicts the factor graph representation of a decomposed joint probability $\mathrm{P}(\mathrm{x},\mathrm{y}) = \mathrm{P}(\mathrm{x} \mid \mathrm{y})\mathrm{P}(\mathrm{y})$, where factor nodes and variable nodes are denoted by squares and circles, respectively. The factor nodes F_1 and F_2 capture the prior probability $\mathrm{P}(\mathrm{y})$ and the conditional probability $\mathrm{P}(\mathrm{x} \mid \mathrm{y})$, respectively. The variable nodes V_1 and V_2 represent the hidden variable x and the observation y, respectively.

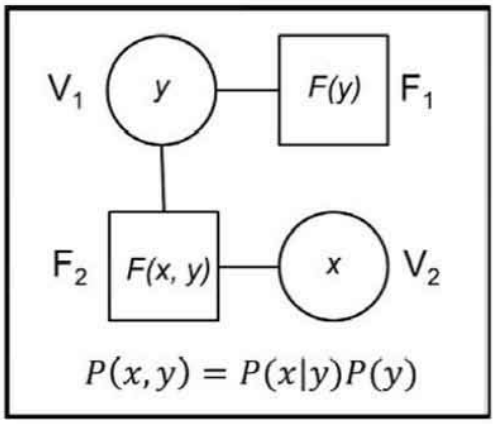


Figure 1. An example of a factor graph representation of a joint probability p(x, y), where factor nodes and variable nodes are denoted by squares and circles, respectively. In this pilot study, we are exploring a factor graph model to represent our knowledge about a corpus of articles related to Kawasaki disease (Kawasaki corpus).

METHODS

Dataset preparation

To evaluate the performance of the factor graph model, we created a Kawasaki disease corpus of 770 PMC articles with automatically extracted references, authors, journal titles, and PMID information. From this corpus, we randomly selected 20 original research articles and manually verified information of MHs, references, and MHs of reference articles. In this model, every article and its reference articles form a factor graph; two graphs are independent from each other. Thus, we can run multiple models in parallel to deal with larger datasets.

Knowledge representation

We used variable nodes to represent PubMed articles; every node has pre-defined attributes to store the knowledge about that article, restricted to the belief distribution on MHs in this report. Variable nodes were connected via intermediary factor nodes according to their citation relations. These intermediary factor nodes summarize beliefs of MHs from adjacent variable nodes and then pass this information to neighbors and update their beliefs on the same MHs. A leaf factor node provided the prior belief distribution on MHs of the connected variable node; MHs that appeared in the corresponding article were assigned greater probabilities while the other MHs received small but non-zero prior probabilities. It is common to see loops in citation networks. In this study, we focused on two-layer tree-structured factor graphs, so we did not have to consider loops. However, it is worth mentioning that the proposed model is able to deal with loops in factor graphs, which refer to the loopy belief propagation algorithm. Although loopy BP will introduce approximation in the results, many existing studies (11) show that loopy BP shows good converge performance. Considering the computation complexity, an intermediary factor node only connected two variable nodes, and a leaf factor node connected one variable node. All cited articles contribute equally to the citing article, given the graph structure and the inference algorithm introduced below.

Inference

We used the sum-product algorithm (a.k.a, belief propagation algorithm) (12,13) to infer the marginal probability on every MH. The sum-product algorithm is an efficient method to compute the exact marginal probability of each variable in an acyclic graph. The sum-product algorithm converges efficiently with acyclic graphs; for graph with cycles, it also provides a good performance in many applications such as image processing. In general, the sum-product algorithm includes three steps.

Step 1: Belief update about each variable: The belief about each variable is the estimated marginal probability of the given variable. In the case of a tree structured graph, the belief is identical to the exact marginal probability once the algorithm converges. The belief update follows equation [1].

$$b(x_i) = \prod_{F_j \in \mathcal{N}(V_i)} m_{F_j \to V_i}(x_i) \qquad [1]$$

where i, j stand for the ith and jth nodes in the factor graph; x_i is the variable represented by the variable node V_i; $\mathcal{N}(V_i)$ is a set of all neighboring factor nodes of V_i; $m_{F_j \to V_i}(x_i)$ is the message sent from adjacent factor node F_j to variable node V_i.

Step 2: Variable node update: The message that will be sent from variable node V_i to an adjacent factor node F_j can be calculated as [2]

$$m_{V_i \to F_j}(x_i) = \frac{b(x_i)}{m_{F_j \to V_i}(x_i)} \qquad [2]$$

Step 3: Factor node update: The message that needs to be sent from factor node F_j to it neighbor variable node V_i can be evaluated as [3]

$$m_{F_j \to V_i}(x_i) = \sum_{V_k \in \mathcal{N}(F_j) \setminus V_i} f(\boldsymbol{x}_s) \prod_{V_k \in \mathcal{N}(F_j) \setminus V_i} m_{V_k \to F_j(x_i)} \quad [3]$$

Where $\mathcal{N}(F_j)$ is a set of all neighboring variable nodes of F_j; $\boldsymbol{x}_s$ is a set of variables represented by the variable nodes in $\mathcal{N}(F_j)$; $f(\boldsymbol{x}_s)$ is the factor function; $\mathcal{N}(F_j) \setminus V_i$ denotes a set of all the neighboring nodes excluding node V_i.

These three steps are repeated until the beliefs converge.

Design of the factor function

The performance of the model largely depends on the design of factor functions. Here, we only consider the MeSH ontology based semantic correlations for estimating the final marginal probabilities. The factor function is a monotonically decreasing function of the semantic correlations. We experimented with different functions from five families: exponential, tangent, arctangent, logarithm, and linear.

Evaluation

We compared the prediction to gold standards in terms of precision, recall, and Mean Average Precision (MAP), which was the mean of the precision scores obtained after each relevant document was retrieved (14–16). The five metrics are defined as below:

$$\text{Precision} = \frac{\sum_D c(N, D, H_1^N)}{\sum_D N}, \quad \text{Recall} = \frac{\sum_D c(N, D, H_1^N)}{\sum_D AN(D)}, \quad \text{F-score} = \frac{2 * precision * recall}{precision + recall}$$

$$\text{AP(D)} = \frac{1}{AN(D)} \sum_r I(h_r) * \frac{c(r, D, H_1^r)}{r}, \quad \text{MAP}(\Omega) = \frac{1}{|\Omega|} \sum_{D \in \Omega} AP(D)$$

H_1^N is a ranked list of top *N* MHs from factor graphs; $c(N, D, H_1^N)$ is the number of correct predictions among the top *N* MHs in document *D*; $AN(D)$ is the number of MHs assigned to *D* in gold standards; $AP(D)$ is the average precision; $I(h_r)$ is an indicator function, which returns 1 if r^{th} MH in the prediction is in the gold standards and return 0 otherwise; Ω is the corpus of articles (15).

We implemented the above evaluation metrics. In addition, we used TREC_EVAL package version 9.0 (16) to calculate MAP. In this study, we had no plan to learn an optimal factor function, because we focused on the possibility of applying factor graph models to MH assignment.

RESULTS

Our two-layer factor graph model is illustrated in Figure 2.

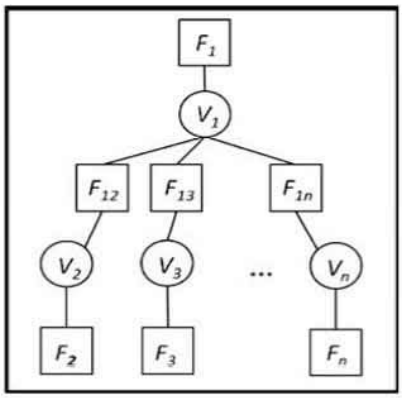


Figure 2. An example of a factor graph with two layers of variable nodes. V_1 is a variable node which represents the citing article. V_2 to V_n are the variable nodes for cited articles. F_1 to F_n are leaf factor nodes and they provide the prior probability distributions on MHs for their adjacent variable nodes. F_{12} to F_{1n} are intermediary factor nodes.

To determine a factor function in this study, we explored five factor functions (Table 1) from four different families and evaluated their MAP scores on five manually verified articles.

Factor Functions
$f(x)=\exp(-x^2)$, $f(x)=\exp(-x^3)$,
$f(x)=-\ln(x)$, $f(x)=-\frac{1}{x}$, $f(x)=\arctan(x)$

Table 1. Candidate factor functions. They are basic functions selected from four families. Variable x is the semantic distance between two MHs in the MeSH ontology.

Five articles were selected from the Kawasaki corpus as shown in Table 2. We manually verified all the selected articles and corrected metadata and reference information collected from automated extraction. On every article, we built factor graph models with five factor functions and evaluated the MAP scores using Trec_eval 9.0 package. Since the exponential functions generally performed better than other functions in this testing, we selected $\exp(-x^2)$ with considerations of further extension. In future work we will consider learning factor functions from data, such as using an iterative log-linear regression method.

	Citing Article PMID	MAP*	Selected Factor Function
1	11953819	0.6266	$\exp(-x^3)$
2	15611788	0.2487	$\exp(-x^2)$
3	11875736	0.3571	$\exp(-x^2)$
4	16202147	0.8889	$\exp(-x^2)$
5	9874566	1	$\arctan(x)$
6	9874566	1	$-\ln(x)$

***These MAP scores were calculated using Trec_eval 9.0 package.**

Each input file contains only one article with all its available MHs.

Table 2. Model performance with different factor graphs. The last column is the factor function used, among six candidates.

On a set of 20 verified articles, we obtained precision, recall, F score, and average precision (AP) in Table 3.

PMID	Precision	Recall	F score	AP
9874566	0.44	0.52	0.48	0.05
11875736	0.20	0.29	0.24	0.29
11953819	0.56	0.78	0.65	0.28
12556969	0.24	0.33	0.28	0.22
12671708	0.36	0.75	0.49	0.33
12823849	0.12	0.60	0.20	0.40
14676801	0.40	0.63	0.49	0.25
15611788	0.20	0.63	0.30	0.63
15928668	0.44	0.53	0.48	0.24
16202147	0.56	0.58	0.58	0.33
16404364	0.52	0.65	0.58	0.55
16594731	0.20	0.38	0.26	0.23
16965625	0.56	0.78	0.65	0.44
17640353	0.24	0.32	0.28	0.21
18070342	0.40	0.56	0.47	0.11
18171482	0.52	0.73	0.60	0.50
18387181	0.44	0.69	0.54	0.38
18782781	0.60	0.79	0.68	0.05
19065999	0.16	0.29	0.21	0.07
19264792	0.52	0.76	0.62	0.47

Table 3. Outcomes from 20 articles.

Based on Table 3, we evaluated the mean Precision, mean Recall, mean F score and MAP as shown in Table 4, following the formula in the evaluation section.

Precision	Recall	F score	MAP
0.38	0.58	0.46	0.30

Table 4. Mean values of precision, recall, F score and average precision.

DISCUSSION

Huang et al. (15) reported a precision of 0.302, recall of 0.583, F score of 0.398, and MAP of 0.462 of MTI on a dataset of 1000 randomly selected MEDLINE documents. We learned from the NLM that the estimated MAP of MTI is 0.35. Considering that MTI adopted multiple attributes such as nearest neighbors and text from titles and abstracts, the factor graph model has shown the potential for providing better ranked MH suggestions.

The performance of the factor graph model depends on multiple factors, including the design of the factor function, the attributes, the type of article, number of references, and the MeSH vocabulary of a particular corpus. In future studies, we will extend the model and incorporate attributes of journal and author, because a journal has a strong association with the topics of its articles and authors usually have

very specific research fields. Other attributes are also in consideration, as long as it can better represent the relations between articles and improve the model performance.

Some MHs do not occur in the cited articles, but they could be derived from the text of the citing article. We will use MetaMap to map text in the citing articles to UMLS terms, and identify potential MHs from these UMLS terms, a solution similar to the one used in MTI. This also shed a light on resolving a limitation of this study. Currently, the factor graph models are with articles in PMC with complete references. However, not all articles are indexed by PMC and references may be incomplete. In the case that reference articles containing desired MHs are missed from data, it is possible to recover these MHs using the above natural language processing techniques.

Conclusion

In this pilot study, we experimented with the factor graph model and sum-product algorithm to infer MHs on a Kawasaki disease corpus from PubMed. The results warrant the further investigation using this technique to improve the prediction performance.

ACKNOLEDGEMENT

This work is supported by NIH grant (U54HL108460), NLM (R00LM011392, R21LM012060) and NHGRI (K99HG008175), in part, by the Medical Informatics Training Program at the Lister Hill National Center for Biomedical Communications, NLM.

REFERENCE

1. Zweigenbaum P, Demner-Fushman D, Yu H, Cohen KB. Frontiers of biomedical text mining: current progress. Brief Bioinform. 2007 Sep 1;8(5):358–75.

2. Understanding the Vocabulary. U.S. National Library of Medicine. Available from: http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_010.html

3. Mork JG, Yepes AJJ, Aronson AR. The NLM Medical Text Indexer System for Indexing Biomedical Literature. Proceedings of the first Workshop on Bio-Medical Semantic Indexing and Question Answering, a Post-Conference Workshop of Conference and Labs of the Evaluation Forum 2013. Valencia, Spain; 2013.

4. Mork JG, Demner-Fushman D, Schmidt SC, Aronson AR. Recent enhancements to the NLM medical text indexer. Working Notes for CLEF 2014 Conference. Sheffield, UK; 2014.

5. What BioASQ Is About [Internet]. Available from: http://www.bioasq.org/project/about

6. Mao Y, Wei C-H, Lu Z. NCBI at the 2014 BioASQ challenge task: large-scale biomedical semantic indexing and question answering. Proceedings of Question Answering Lab at CLEF. 2014.

7. Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, et al. The NLM Indexing Initiative. Proceedings of the AMIA Symposium American Medical Informatics Association. 2000. p. 17–21.

8. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics. 2009 Jun 1;25(11):1412–8.

9. Liu K, Wu J, Peng S, Zhai C, Zhu S. The Fudan-UIUC Participation in the BioASQ Challenge Task 2a: The Antinomyra system. CLEF (Working Notes). 2014. p. 1311–8.

10. Papanikolaou Y, Dimitriadis D, Tsoumakas G, Laliotis M, Markantonatos N, Vlahavas IP. Ensemble Approaches for Large-Scale Multi-Label Classification and Question Answering in Biomedicine. CLEF (Working Notes). 2014. p. 1348–60.

11. Ihler AT, Fisher JWI, Willsky AS. Loopy Belief Propagation: Convergence and Effects of Message Errors. Journal of Machine Learning Research. 2005. p. 905–36.

12. Kschischang FR, Frey BJ, Loeliger H-A. Factor graphs and the sum-product algorithm. IEEE Trans Inf Theory [Internet]. IEEE Press; 2001;47(2):498–519.

13. Yedidia JS, Freeman WT, Weiss Y. Understanding belief propagation and its generalizations. Morgan Kaufmann Publishers Inc.; 2003 Jan 1;239–69.

14. Buckley C, Voorhees EM. Retrieval evaluation with incomplete information. Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR '04 [Internet]. New York, New York, USA: ACM Press; 2004. p. 25.

15. Huang M, Névéol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. J Am Med Inform Assoc;18(5):660–7.

16. Trec_eval09 package [Internet]. [cited 2014 Sep 3]. Available from: http://trec.nist.gov/trec_eval/

A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports

Alexandre Yahi[1] and Nicholas P. Tatonetti, PhD[1]

[1]Department of Biomedical Informatics, Department of Systems Biology, Department of Medicine, Columbia University, New York, NY, USA

Abstract

The secondary use of electronic health records (EHR) represents unprecedented opportunities for biomedical discovery. Central to this goal is, EHR-phenotyping, also known as cohort identification, which remains a significant challenge. Complex phenotypes often require multivariate and multi-scale analyses, ultimately leading to manually created phenotype definitions. We present Ontology-driven Reports-based Phenotyping from Unique Signatures (ORPheUS), an automated approach to EHR-phenotyping. To do this we identify unique signatures of abnormal clinical pathology reports that correspond to pre-defined medical terms from biomedical ontologies. By using only the clinical pathology, or "lab", reports we are able to mitigate clinical biases enabling researchers to explore other dimensions of the EHR. We used ORPheUS to generate signatures for 858 diseases and validated against reference cohorts for Type 2 Diabetes Mellitus (T2DM) and Atrial Fibrillation (AF). Our results suggest that our approach, using solely clinical pathology reports, is an effective as a primary screening tool for automated clinical phenotyping.

Introduction & Background

Electronic health records (EHR) capture an increasing variety and amount of clinical data leading to initiatives that are leveraging this potential for knowledge discovery. From adverse event and medical error detection for patient safety[1,2] to case-control studies[3], those new tools often rely on the researchers' ability to isolate accurate cohorts of patients with a given phenotype. In this context, the term phenotyping has been used to describe automated and manual methods for identifying these patient cohorts in the EHR[4].Advancement of automated phenotyping algorithms is a major roadblock in the field[4]. Several nationwide efforts, such as eMERGE[5] and SHARPn[6], have developed selection algorithms for high-throughput phenotype extractions. Those algorithms often comprise of a series of arithmetic and logical operations that are applied to the clinical data. The data types used in these algorithms are heterogeneous and may vary between institutions necessitating continual re-evaluation[7]. There is an opportunity in phenotyping to apply statistical learning methods, like Association Rule Mining (ARM), for modeling selection algorithms[8] or the use of tensor factorization of medications and diagnoses to identify patients[9]. Other approaches have focused on certain types of clinical data like the diagnoses codes, which often are ICD-9-CM codes. Machine learning techniques trained on these data have been able to classify patients even when data are missing by using inductive logical programming[10]. The exclusive use of a particular clinical data type (e.g., medications or clinical pathology reports) is advantageous because it allows the exploration other the other data types in the selected cohort while minimizing bias to the extent possible. In particular, ICD-9-CM codes have been widely used for phenotyping and, in some cases, enhanced by additional information, such patient-reported data[11]. However, ICD-9-CM are primarily used for billing purposes and not for differential diagnosis, introducing complicated biases[12]. *Clinical pathology* is the medical subfield that deals with the analysis of bodily fluids for diagnosis and prognosis and clinical pathology reports, commonly called "lab reports," may be more reliable than ICD-9 codes for EHR phenotyping, while maintaining the same level of standardization.

We present Ontology-driven Reports-based Phenotyping with Unique Signatures (ORPheUS), a knowledge-based phenotyping method that generates a unique clinical pathology signature for each term of a given ontology (i.e. each disease phenotype). Each "phenotype signature" is comprised of a set of abnormal laboratory tests (ATs). Our approach relies on only one type of clinical data -- the clinical pathology reports -- to minimize biases and increase interoperability. In total we generated clinical pathology signatures for 858 distinct diseases. We validated three of these signatures against reference patient cohorts using definitions from PheKB.org. We evaluated for precision and recall as well as the recovery of known co-morbidities. In each case we found that ORPheUS significantly outperforms the null model, with the T2DM signature recovering 17.2% of diabetics at 81.4% precision (F1 score=0.28).

Methods

Clinical Data Sources

The New York Presbyterian/Columbia University Medical Center (NYP/CUMC) clinical data warehouse contains about 470 million laboratory values from clinical pathology reports from more than 1.3 million patients over the last decade. We selected 177 of the most commonly ordered tests performed from blood, urine, plasma, and cerebrospinal fluid. We restricted our cohort of study to patients over 18 years old at order time with specified sex and at least one of these 177 laboratory tests. It narrowed our study to 767,389 patients with 172,518,869 values total. We preprocessed these data to assert if those reports were normal, abnormal, high, or low accounting for the patients' age and sex, and according to our normal ranges database (Yahi, et al, *in preparation*).

Annotating abnormal laboratory tests with ontology terms

ORPheUS uses abnormal laboratory tests (ATs). We associated each AT to the medical terms from a given ontology through statistical enrichment analysis. We created the initial set of annotations by defining a search term by concatenating the name of the laboratory test with its non-normal status (i.e., "blood glucose low", "blood glucose high", etc.). Then we searched for each of these terms in the medical search engine UpToDate (www.uptodate.com) and gathered the titles of the first three pages of results. Once regrouped in a text file, these titles were annotated with the Annotator API by the NCBO (www.bioontology.org)[13] and counted the number of times an ontology term would appear. We attributed 10 points for an exact match and 8 points for a synonym match. . This is a one-time process associate ATs to clinical ontology terms and it is not repeated for the following steps of the phenotyping. We looped through all the terms of the ontology to associate each medical term with the ATs associated with its semantic descendants. We performed a Fisher's exact test and a permutation analysis on these annotations sets to identify the ATs significantly associated to each ontology term, assessing significance using a FDR <= 0.05. Therefore, each ontology term (e.g., "Diabetes mellitus"), we have a set of significant ATs. We call this set of ATs the phenotype signature.

Selecting cohorts of patients for reference standard

We applied phenotype selection algorithms available on PheKB (www.phekb.org) to construct a reference standard. We therefore identified case cohorts for Atrial Fibrillation (AF)14 and Type 2 Diabetes Mellitus (T2DM)[15,16]. The data required by these algorithms consists of ICD-9 codes, CPT-4 codes, drug prescriptions, and clinical notes. We tested the performance of ORPheUS on these reference groups of patients.

Phenotyping with ORPheUS

We identified the presence of the phenotype signatures, complete (i.e., all the ATs of the signature are found in the patient's clinical history) or partial (i.e. a subset of the ATs in the signature), in a patient's clinical pathology records. For each patient, we look for the presence of any of the ATs belonging to the signature in his medical record to consider this patient as a potential candidate. We referenced laboratory tests with a universal code system named Logical Observation Identifiers Names and Codes (LOINC)[17] and we used these codes to match ATs. We sorted those candidates by the number of distinct ATs of the target signature they had without any constraint in time. We designated by true positive (TP) the patients at the intersection of each of these prediction sets and its reference cohort of patients. To assess statistical significance, we compared the precision of the predictions from the signatures to a randomly selected cohort of the same size. For each group of candidates with N distinct ATs, we compared the precision of the prediction against the precision of a randomly selected cohort of the same size relative to all the patients with at least N distinct clinical pathology reports. We performed this random selection 20 times for each category. To compute the recall, we proceeded the same way except that the predictions were evaluated against the complete cohort of reference patients.

Results

Signatures

We annotated 351 abnormal laboratory test (ATs) with terms from the Human Disease Ontology (DOID)[18]. We then identified those ATs that were specific to each term to generate 858 signatures. The average signature contained 10.8 ± 14 ATs. The minimum number of ATs in a signature was 1 (for 95 signatures), and the maximum 50 (DOID:1579 Respiratory system disease). We did not construct a signature for parent term, "Disease," in the ontology. Diabetes Mellitus with 14 distinct ATs is a little above the average of signatures (Table 1 – Signature for Diabetes Mellitus). Congenital heart disease presents 16 ATs and Mycoardial infarction 14 (Table 2 and 3).

Diabetes Mellitus (DOID:9351)

Clinical Pathology Report	Status
Glucose in Serum or Plasma	High/Low
Fasting glucose in Serum or Plasma	High/Low
Glucose in Blood	High/Low
Glucose in Serum or Plasma post challenge	High/Low
Hemoglobin A1c/Hemoglobin.total in Blood by HPLC	High/Low
Glucose in Blood (Meter)	High/Low
Hemoglobin A1c/Hemoglobin.total in Blood	Low
Hemoglobin in Blood	High

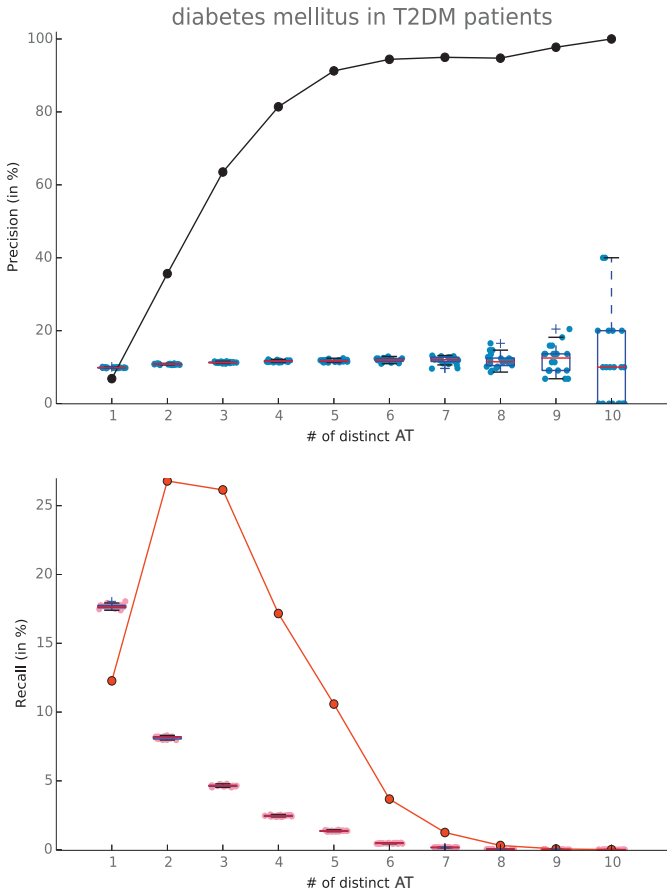
Table 1 – Signature of Diabetes Mellitus (DOID:9351)

congenital heart disease (DOID:1682)

Clinical Pathology Report	Status
Carbon dioxide, total in Arterial blood	High/Low
Carbon dioxide, total in Serum or Plasma	High
Estradiol (E2) in Serum or Plasma	High
Thyroxine (T4) free in Serum or Plasma	High
Calcium.ionized in Arterial blood	High
Erythrocyte mean corpuscular volume by Automated count	Low
Oxygen saturation in Arterial blood	High/Low
Oxygen saturation Calculated from oxygen partial pressure in Blood	High
Oxygen saturation in Venous blood	High/Low
Oxygen [Partial pressure] in Arterial blood	High/Low
Oxygen [Partial pressure] in Venous blood	Low
Thyroxine (T4) in Serum or Plasma	High

Table 2 – Signature Congenital heart disease (DOID:1682)

Figure 1. (left) Precision and Recall curves for Diabetes Mellitus signatures tested on T2DM patients



myocardial infarction (DOID:5844)

Clinical Pathology Report	Status	Clinical Pathology Report	Status
Basophils [#/volume] in Blood	High	Platelet mean volume in Blood	High
Eosinophil [#/volume] in Blood	High	INR in Platelet poor plasma by Coagulation assay	High
Eosinophils [#/volume] in Blood by Manual count	High	Carbon dioxide [Partial pressure] in Arterial blood	High
Fibrinogen in Platelet poor plasma by Coagulation assay	High	Platelets in Blood	High
Hematocrit of Blood by Automated count	High	Potassium in Arterial blood	High
Hematocrit of Blood	Low	Sirolimus in Blood	High
International Normalized Ratio POC	High	Thrombin time in Platelet poor plasma by Coagulation assay	High

Table 3 – Signature of myocardial infarction (DOID:5844)

Phenotyping performances

We computed the precision and recall curves for the Diabetes Mellitus in 83,246 patients with T2DM as determined by the reference standard. We observed that of the 14 T2DM specific ATs in the signature, we only found up to 10 simultaneously in a single patient's record. The precision is significantly better than by chance and increases above 80% with when at least 4 ATs are matched. At 6 or more distinct ATs the recall falls to below 5% (Figure 1).

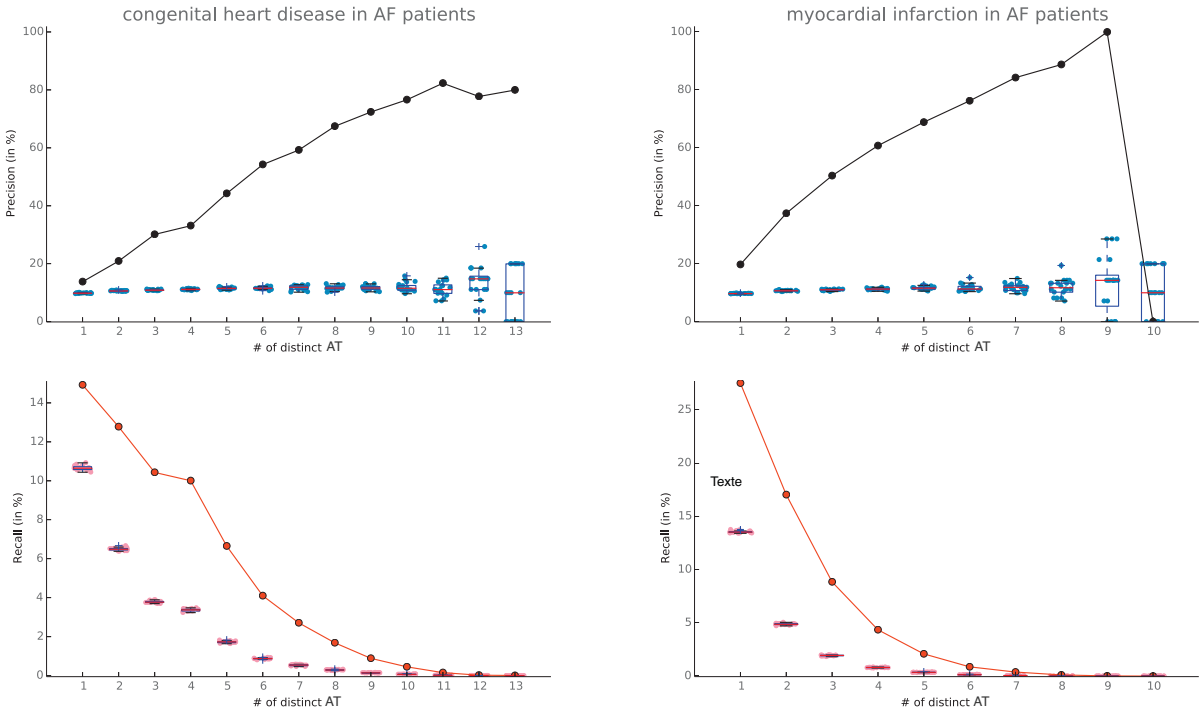


Figure 2. (a) Congenital heart disease and (b) Myocardial infarction signatures in Atrial Fibrillation patients

We also explored the cohorts of 80,163 patients with Atrial Fibrillation and evaluated the signatures of two of AF's known comorbidities: myocardial infarction[19], and congenital heart disease[20]. We observed an interesting precision for Congenital Heart Disease (Figure 2.a.) reaching a plateau around 80% from 10 distinct ATs. Myocardial Infarction (Figure 2.b.) presented a better precision, needing only 6 distinct ATs to reach 80%. However, despite a better initial recall, we witnessed a faster drop in sensitivity for the myocardial infarction signature than the congenital heart disease one. Finally, we observed that for 10 distinct ATs the predicted set of patients was so small that the precision fell to zero.

Discussions

In this paper we present a novel automated EHR phenotyping algorithm by defining signatures of abnormal laboratory tests and scanning for matches in a patient's longitudinal medical record. These signatures are knowledge-driven and rely on only one type of clinical data helping to minimize biases and improve interoperability. Since the signatures are knowledge-based they are not directly exposed to any clinical data before they are used for phenotyping. In total we generated 858 disease signatures. We validated two (atrial fibrillation and type 2 diabetes mellitus) of these signatures against a reference cohort of patients identified using eMERGE algorithms available at PheKB.org. We did not revalidate the PheKB algorithms in the CUMC database, however, previous implementations showed a 98% Positive Predictive Value for AF, and between 98 and 100% for T2DM.

In future studies, we would like to consider co-occurrences of those signatures across time. We might consider restricting the time windows from 1 to 12 months in patients' records and look for the phenotype signatures, keeping only the maximum number of distinct simultaneous ATs in these windows. It might improve the precision of our predictions since some patients present sparse clinical pathology reports. Dynamical phenotyping using those reports has shown promising opportunities[21]. We would also like to investigate the potential of combining different phenotypes signatures. We also envision a possible approach for robustness assessment, which would consist of

mapping ontological terms, in this example a DOID term, to ICD-9-CM diagnoses codes. This would allow us to evaluate performance of our all or most of our generated phenotype signatures systematically.

The EHR systems are in constant evolution, and many efforts are focused on designed new models learning from data and mitigate complex, inaccurate and frequently missing clinical values[4]. Indeed, the need for normalization in the information models that are use and the use of standardized vocabularies would ensure a better end-to-end connectivity over platforms allowing more reliable high-throughput phenotyping[6]. Meanwhile, as clinical notes still remain a critical source of information for phenotypic characteristics, phenotyping techniques using natural language processing (NLP) has been widely used and are gaining popularity[22]. The term of "Verotype" as a matching of genotype, phenotype and disease subtype has also been described[23] to make a step forward to personalized medicine. The systematic inclusion of genotype and phenotype data in future EHR would be critical for this purpose[24].

Conclusion

We presented Ontology-driven Reports-based Phenotyping with Unique Signatures (ORPheUS), a knowledge-based automated method for EHR-phenotyping, using only clinical pathology reports. We evaluated the performances of our phenotype signatures for T2DM and AF and demonstrated the potential use of this method for phenotyping. Our ontology-driven approach could allow us in future work to use other medical semantic fields and study for example adverse events signatures.

References

1. Li Q, Melton K, Lingren T, Kirkendall ES, Hall E, Zhai H, et al. Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2014 Sep;21(5):776–84.
2. Lorberbaum T, Nasir M, Keiser MJ, Vilar S, Hripcsak G, Tatonetti NP. Systems pharmacology augments drug safety surveillance. Clin Pharmacol Ther. 2014 Nov 1.
3. Castro VM, Mahamaneerat W, Gainer VS, Ananthakrishnan AN, Porter AJ, Wang TD, et al. Evaluation of matched control algorithms in EHR-based phenotyping studies: A case study of inflammatory bowel disease comorbidities. J Biomed Inform. 2014 Sep 6.
4. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2013 Jan 1;20(1):117–21.
5. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, et al. Electronic medical records for genetic research: results of the eMERGE consortium. Science Translational Medicine. American Association for the Advancement of Science; 2011 Apr 20;3(79):79re1–79re1.
6. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2013 Dec;20(e2):e341–8.
7. Overby CL, Weng C, Haerian K, Perotte A, Friedman C, Hripcsak G. Evaluation considerations for EHR-based phenotyping algorithms: A case study for drug-induced liver injury. AMIA Jt Summits Transl Sci Proc. 2013;2013:130–4.
8. Li D, Simon G, Chute CG, Pathak J. Using association rule mining for phenotype extraction from electronic health records. AMIA Jt Summits Transl Sci Proc. 2013;2013:142–6.
9. Ho JC, Ghosh J, Steinhubl S, Stewart W, Denny JC, Malin BA, et al. Limestone: High-throughput candidate phenotype generation via tensor factorization. J Biomed Inform. 2014 Jul 16.
10. Peissig PL, Santos Costa V, Caldwell MD, Rottscheit C, Berg RL, Mendonca EA, et al. Relational machine learning for electronic health record-driven phenotyping. J Biomed Inform. 2014 Jul 15.
11. Griffith SD, Thompson NR, Rathore JS, Jehi LE, Tesar GE, Katzan IL. Incorporating patient-reported outcome measures into the electronic health record for research: application using the Patient Health Questionnaire (PHQ-9). Qual Life Res. Springer International Publishing; 2014 Aug 7;:1–9.
12. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. J Biomed Discov Collab. 2011;6(0):48–52.
13. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucl Acids Res. Oxford University Press; 2009 Jul;37(Web Server issue):W170–3.
14. Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust Replication of Genotype-Phenotype Associations across Multiple Diseases in an Electronic Medical Record. The American Journal of Human Genetics. 2010 Apr;86(4):560–72.
15. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2012 Mar;19(2):212–8.
16. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2012 Mar;19(2):219–24.
17. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clinical Chemistry. 2003 Apr;49(4):624–33.
18. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucl Acids Res. Oxford University Press; 2012 Jan;40(Database issue):D940–6.
19. Soliman EZ, Safford MM, Muntner P, Khodneva Y, Dawood FZ, Zakai NA, et al. Atrial Fibrillation and the Risk of Myocardial Infarction. JAMA Intern Med. American Medical Association; 2014 Jan 1;174(1):107–14.
20. Andrade J, Khairy P, Dobrev D, Nattel S. The clinical profile and pathophysiology of atrial fibrillation: relationships among clinical features, epidemiology, and mechanisms. Circulation Research. Lippincott Williams & Wilkins; 2014 Apr 25;114(9):1453–68.
21. Albers DJ, Elhadad N, Tabak E, Perotte A, Hripcsak G. Dynamical phenotyping: using temporal analysis of clinically collected physiologic data to stratify populations. Garcia-Ojalvo J, editor. PLoS ONE. Public Library of Science; 2014;9(6):e96443.
22. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2014 Mar;21(2):221–30.
23. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. Journal of the American Medical Informatics Association. BMJ Publishing Group Ltd; 2013 Dec;20(e2):e232–8.
24. Frey LJ, Lenert L, Lopez-Campos G. EHR Big Data Deep Phenotyping. Contribution of the IMIA Genomic Medicine Working Group. Yearb Med Inform. 2014;9(1):206–11.

Mining Biomedical Literature to Explore Interactions between Cancer Drugs and Dietary Supplements

Rui Zhang, PhD[1, 2], Terrance J. Adam, RPh, MD, PhD[1, 3], Gyorgy Simon, PhD[1], Michael J. Cairelli, DO, MS[4], Thomas Rindflesch, PhD[4], Serguei Pakhomov, PhD[1, 3], Genevieve B. Melton, MD, MA[1, 2]
[1]Institute for Health Informatics; [2]Department of Surgery; [3]College of Pharmacy, University of Minnesota, Minneapolis, MN; [4]Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health

Abstract

Interactions between cancer drugs and dietary supplements are clinically important and have not been extensively investigated through mining of the biomedical literature. We report on a previously introduced method now enhanced by machine learning-based filtering. Potential interactions are extracted by using relationships in the form of semantic predications. Semantic predications stored in SemMedDB, a database of structured knowledge generated from MEDLINE, were filtered and connected by two interaction pathways to explore potential drug-supplement interactions (DSIs). The lasso regression filter was trained by using SemRep output features in an expert annotated corpus and used to rank retrieved predications by predicted precision. We found not only known interactions but also inferred several unknown potential DSIs by appropriate filtering and linking of semantic predications.

Introduction

The use of natural supplements in the US has increased dramatically in recent years. According to the results of a National Health Interview Survey in 2012, 17.9% of American adults had used dietary supplements (not including vitamins and minerals).[1] National Health and Nutrition Examination Survey (NHANEs) data also indicated that 53% of American adults took at least one dietary supplement during 2003-2006, mostly multivitamin and multimineral supplements.[2] When taking occasional and seasonal use into account, the prevalence of supplement use was 69% in 2011. Additionally, supplement use in women is higher than in men according to consumer surveys by the Council for Responsible Nutrition.[3] Those who use the supplements often take them in combination with conventional drugs. About 30% of the elderly population (age >65), the largest group consuming prescription drugs, use at least one daily supplement, thus placing the patient at risk for potential drug-supplement interactions (DSIs). Supplements are also increasingly used in the US by patients diagnosed with cancer to help strengthen their immune system and ease the side effects of treatments.

The growing popularity of supplements has focused attention on DSIs.[4] One study suggested that patients on medications with a narrow therapeutic index (e.g., cyclosporine, phenytoin, warfarin) should avoid the use of herbal products, as those drugs may either have adverse effects or be less effective when combined with such products.[5] Gurley et al. reported that the concomitant administration of botanical supplements with P-glycoprotein (P-gp) substrates can lead to clinically significant interactions. The study was based on evaluating the effects St. John's wort (SJW) and Echinacea on the pharmacokinetics of a P-gp substrate, digoxin.[6] It was suggested that the concomitant use of docetaxel and SJW should be avoided in cancer patients, as the hyperforin component in SJW can induce cytochrome P450 3A4, thus leading to changes in drug metabolism for a number of chemotherapeutic and other conventional drugs.

Many of these studies have focused on limited sets of supplements and drugs. Most supplements have not been studied extensively in clinical trials. Some serious DSIs are not found until a new drug is already on the market, since clinical trials for new drugs do not typically consider DSIs. Therefore, many DSIs are unknown to both health care providers and patients themselves. Current DSI documentation is limited, as it is only based on pharmacological, in vitro, or animal model data. Moreover, because of the less rigorous regulatory rules regarding dietary supplements, formulations may vary significantly by manufacturer, and similar products may be derived from a variety of sources.

In addition, potential DSIs may result from undefined pathways that have yet to be discovered. Such interactions often can only be derived with an indirect approach, such as mining the scientific literature. This resource contains a large amount of pharmacokinetic and pharmacodynamic knowledge in free text and expands the range of drugs, supplements and genes. Compared to traditional drug-drug interaction work, the use of literature-based discovery for DSI identification has not been adequately investigated. We hypothesize that a powerful literature-based information discovery system could significantly enhance DSI knowledge bases and further translate to clinical practice for

increased quality of patient care. In this study, we investigated the use of the structured knowledge extracted from biomedical literature for exploration of DSIs.

Background

Literature-based discovery (LBD) is an automatic method to generate hypotheses by connecting findings in the literature. In general, if a concept Y is related to both concept X and concept Z, there exists a potential association between X and Z. For example, Swanson et al. applied this approach to propose fish oil as a potential treatment for Raynaud's disease.[7] Hristovski et al. proposed to enhance LBD by using semantic predications instead of depending on co-occurrence of words or concepts.[8] They found that linking semantic predications could generate a larger number of useful findings. In this study, we will treat a gene as the concept Y linking a cancer drug (X) and a dietary supplement (Z).

SemRep is a semantic interpreter that extracts structured knowledge in the form of semantic predications from MEDLINE citations[9]. Each predication consists of two UMLS Metathesaurus concepts as subject and object, and a semantic relationship (from the UMLS Semantic Network) as a predicate. SemMedDB[10] is a database containing semantic predications generated by SemRep from over 23.6 million MEDLINE citations. The database for this study contains 69 million semantic predications from processed citations published as of March 31, 2014. SemRep and SemMedDB have been used in pharmacogenomic information extraction[11, 12] and information extraction for clinicians' needs[13].

Methods

The overall methodology includes 1) drug and supplement concept mapping, 2) related semantic predication extraction, 3) machine-learning based filtering, 4) extraction of potential interaction pathways, and 5) expert judgment and verification of known interactions.

Supplements and cancer drugs

In this study, we focused on a limited subset of dietary supplements and cancer drugs. A supplement list was obtained from the Office of Dietary Supplements website (http://ods.od.nih.gov/factsheets/list-all/ - accessed August 1, 2014). Drugs approved by the FDA to treat breast and prostate cancer were also collected from the National Cancer Institute website (http://www.cancer.gov/cancertopics/druginfo/drug-page-index - accessed August 1, 2014). The corresponding CUIs for supplements and cancer drugs were retrieved by mapping the supplements and drugs to the UMLS Metathesaurus via MetaMap.

Extraction of semantic predications

We extracted from SemMedDB predications with various semantic types: supplement-gene (i.e., predications with a dietary supplement as the subject and a gene as the object), breast_cancer_drug-gene, prostate_cancer_drug-gene, gene-supplement, gene-breast_cancer_drug, and gene-prostate_cancer_drug. We restricted our analysis to three predicate types: STIMULATES, INHIBITS, and INTERACTS_WITH.

Machine learning-based filter

Considering the limited precision of SemRep that we have seen previously[11], we developed a machine learning (ML)-based filter to rank the generated semantic predications by probability of being correct. We used the reference standard from a previous study containing annotations for 300 randomly selected sentences that involve drug-gene, gene-drug and gene-biological function predications (dataset details described in the paper)[11] as our training set. SemRep extracted 524 total semantic predications from these sentences, 304 of which were evaluated to be correct. In this study, we only used predications generated by SemRep that were judged by experts as either true positive (TP) or false positive (FP). The supervised machine-learning algorithm, lasso regression (LR), was used to classify SemRep output as either TP or FP. We used SemRep output features as predictors, including UMLS biomedical concepts, argument distance, indicator types (e.g., verb), predicate (e.g., TREATS), and UMLS semantic types.

To evaluate the effectiveness of the filter, we ranked the generated semantic predications based on the score the model assigned to each semantic predication. A physician (MJC) was asked to judge the top 20 for each type of predication (including supplement-gene, gene-supplement, breast cancer drugs-gene, gene-breast cancer drugs, prostate cancer drugs-gene, gene-prostate cancer drugs). The precision was calculated as *True Positives/(True Positives+False Positives)*.

DSI discovery

We further filtered out some nonspecific concepts such as "gene", "receptor" and "proteins" before discovery. We ranked the potential DSIs based on two pathways modified from earlier work[11]: Drug→Gene→Supplement and Supplement→Gene→Drug schemas. For the first pathway, when semantic predications Drug→Gene and

Gene→Supplement share the same gene, an interaction may be indicated. For example, the predications Echinacea INHIBITS CYP450, and CYP450 INTERACTS_WITH Docetaxel generate the potential interaction Echinacea→CYP450→Docetaxel. A score was then assigned to each potential interaction (e.g., Supplement→Gene→Drug) by adding two rank scores of Supplement→Gene and Gene→Drug, which were obtained from the ML-based filter. These interactions were ranked based on this score and then evaluated.

Expert selection and database checking

A drug interaction expert that is a pharmacist and physician (TJA) manually reviewed the top 100 ranked interactions. Priority for review of potential interactions was first given to gene specificity, with priority for specific gene paths (e.g. CYP 450 3A4) preferred over more general gene categories (CYP p450). Second priority was given to highly plausible combinations such as those including supplements to enhance the immune system or help with chemotherapy side effects, where the supplements would be likely to be used in combination in a clinical setting. The next priority was the level of evidence in describing potential interactions and was assessed based on the pharmacologic data, with priority given to interaction data providing evidence of substance-gene activity.

To compare potential DSIs found by our method with established interactions, a drug profile for drugs for breast and prostate cancer, as well as a number of targeted supplements was entered into a well-known drug-drug interaction (DDI) website (http://cpref.goldstandard.com/interreport.asp) with a theoretical multiple drug profile including pertinent supplements and chemotherapy drugs. Another website, at Case Western Reserve University Hospital, was used to provide additional support: http://www.uhhospitals.org/health-and-wellness/drug-information-center/drug-interaction-tool.

Results

We extracted 10,500 supplement→gene predications, 270 prostate_cancer_drug→gene predications, 991 breast_cancer_drug→gene predications, 7732 gene→supplement predications, 280 gene→breast_cancer_drugs predications, and 217 gene→prostate_cancer_drugs predications. The precision of the top ranked predications after filtering was 69%, a significant increase over the 58% previously reported for a randomly selected set of similar predications[10]. After combining the top ranking 500 predications from each side of the pathway (supplement-gene and gene-drug), 1095 combinations were formed and ranked by score. After expert review, we examined some of these DSIs focusing on the pharmacologically active CYP450 gene family, which has potential effects on multiple therapeutic classes. We found both known and unknown DSIs, and we found five more interactions with filtered predications than with unfiltered predications. Some of the potential DSIs identified shared the same pathway, such as the interactions between Echinacea and chemotherapeutic medications cyclophosphamide, docetaxcel, everolimus, fluorouracil and toremifine (Table 1). Table 2 lists selected semantic predications and citations.

Table 1. Selected DSI examples and pathways. INH, INHIBITS; STI, STIMULATES; INT, INTERACTS_WITH.

Drug/Supplement	Predicate	Gene/Gene Class	Predicate	Supplement/Drug	Known	Filter/ Unfilter
Echinacea	INH	CYP450	INT	Cyclophosphamide	Y	Both
Echinacea	INH	CYP450	INT	Docetaxel	Y	Both
Echinacea	INH	CYP450	INT	Everolimus	Y	Both
Echinacea	INH	CYP450	INT	Fluorouracil	Y	Both
Echinacea	INH	CYP450	INT	Toremifine	N	Both
Echinacea	STI	CYP1A1	INT	Exemestane	N	Both
Grape seed extract	INH	CYP3A4	INT	Docetaxel	N	Both
Kava preparation	STI	CYP3A4	INT	Docetaxel	Y	Filter
Ginseng	INH	CYP3A	INT	Ginkgo bilob extract	Y	Unfilter
Ginseng	INH	CYP3A	INT	Docetaxel	N	Unfilter
Prednisone	INT	P-glycoprotein	STI	Vitamin E	N	Filter
Cyclophosphamide	INT	P-glycoprotein	STI	Vitamin E	N	Filter
Glucosamine	INH	COX2	STI	Docetaxel	Y	Filter
Melatonin	INH	COX2	STI	Docetaxel	N	Filter

Table 2. Selected semantic predications and citations.

Semantic Predications	Citations (PMID)
Echinacea STIMULATES CYP1A1	Our in vivo data indicate that the Echinacea ethanolic extract can potently inhibit the expression of CYP3A1/2 and can also induce of CYP1A1, CYP2D1. (20374973)
Grape seed extract	Four brands of GSE had no effect, while another five produced mild to moderate but variable

INHIBITS CYP3A4	inhibition of CYP3A4, ranging from 6.4% by Country Life GSE to 26.8% by Loma Linda Market brand. (19353999)
Melatonin INHIBITS Cyclooxygenase-2	Moreover, Western blot analysis showed that melatonin inhibited LPS/IFN-gamma-induced expression of COX-2 protein, but not that of constitutive cyclooxygenase. (18078452).
Prednisone INTERACTS_WITH P-glycoprotein	PRED is also a substrate of P-gp and is a weak inducer of CYP3A, and drug-drug interactions within this combination therapy might occur. (23267661)
Cyclophosphamide INTERACTS_WITH P-glycoprotein	These findings suggest that active cyclophosphamide metabolite can be a substrate for P-glycoprotein. (22803083)
CYP450 INTERACTS_WITH Toremifene	Tamoxifen and toremifene are metabolised by the cytochrome p450 enzyme system, and raloxifene is metabolised by glucuronide conjugation. (12648026)
CYP3A INHIBITS Docetaxel	Because docetaxel is inactivated by CYP3A, we studied the effects of the St. John's wort constituent hyperforin on docetaxel metabolism in a human hepatocyte model. (16203790)
CYP1A1 INTERACTS_WITH Exemestane	Recombinant CYP1A1 metabolized exemestane to MI with a catalytic efficiency (Cl(int)) of 150 nl/pmol P450 x min that was at least 3.5-fold higher than those of other P450s investigated. (20876785)
Cyclooxygenase 2 STIMULATES Docetaxel	We investigated whether prostate tumor-associated stromal cells, marrow-derived osteoblasts, affect cytotoxicity of 2 antitumor drugs, COL-3 and docetaxel (TXTR), and whether it is dependent on COX-2 activity. (15688368)
P-glycoprotein STIMULATES Vitamin E	Expression of multiple drug resistant (MDR) phenotype and over-expression of P-glycoprotein (P-gp) in the human hepatocellular carcinoma (HCC) cell clone P1(0.5), derived from the PLC/PRF/5 cell line (P5), are associated with strong resistance to oxidative stress and a significant ($p < 0.01$) increase in intracellular vitamin E content as compared with the parental cell line. (15453640)

Discussion

DSI is an important topic and deserving of additional investigation with informatics methods to explore potential interactions extracted from the biomedical literature that may have a significant effect on medication therapy. In this study, we investigated the use of semantic knowledge provided by SemRep for DSI extraction. However, due to its limited recall and precision, human review is typically required for maximally effective use of this resource. This intensive manual process of filtering has limited the general use of SemRep in larger scale applications in biomedical and health informatics. Although argument-predicate distance has been used to enhance the precision of extracting semantic predications[14], the use of machine-learning for automatic filtering of semantic predications has not been investigated. The ranking score for each potential interaction was used as an additional means lower the number of interactions subjected to human review. It was found that the filter helped to discover not only those found by using unfiltered results but also found several additional DSIs. Two DSIs that were found without filtering were not found in the filtered interactions.

Both known and unknown interactions were found. The first DSI candidate area of interest is Echninacea which has been known to affect chemotherapy drugs. Cyclophosphamide, docetaxel, fluorouracil are all standard therapeutics for breast and prostate cancer patients and all were noted in our data, as well as in the DDI medication site used for confirmation. We also identified a potentially novel DDI with Echinacea. Exemestane was noted to have an interaction with Echinacea in the test data, with specific activity identified at the CYP1A1 gene. This was confirmed after expert review. This interaction did not show up on the DDI sites consulted. This may be worth additional exploration, especially since metabolites of Exemestane result in reduced activity or non-activity of this medication. The potential implication of therapeutic failure may have an impact on patient survival. In Another example involves P-glycoprotein, which can affect many drugs. Prednisone and cyclophosphamide were identified to be substrates for P-glycoprotein with both having the potential for interactions with the other. In addition they both may interact with the supplement Vitamin E. Melatonin inhibits COX2 via suppression of protein expression potentially creating an interaction with docetaxel which also has activity via the COX2 pathway. Ginseng was also explored for possible interaction in our theoretical patient profile. It was noted to interact with ginkgo on the DDI checker website. This is confirmed in our test data set via CYP450 pathway. In the case of prostate cancer, the potential effect of ginseng on multiple CYP450 pathways was noted, which may result in a DSI with docetaxel through the CYP450 3A pathway. This interaction is not noted at either of the DDI websites we consulted and is an area for future exploration.

In this pilot study, we did not define the specific interaction relations between cancer drugs and dietary supplements,

although we found their potential pathways. In previous SemRep error analysis11, false positives of semantic predications were mainly due to knowledge source shortcomings (e.g., incorrect mapping of gene or protein mentions to UMLS concepts or Entrez Gene terms) and SemRep processing shortcomings with linguistic phenomena (e.g., negation, serial coordination). Future SemRep development efforts include addressing these shortcomings.

Although known interactions were discovered using this methodology, the goal was to identify unknown interactions. Verifying such interactions is a significant challenge for this type of methodological development since, by definition, there is no reference standard. The best validation method would be to translate the findings into a clinical trial for the suggested combination. While our approach of identifying literature support for the proposed interaction is significantly more expedient, it is less robust and further clinical evidence is required to fully validate these findings.

In conclusion, we found both known and unknown DSIs by using combining a supplement→gene→drug schema data with LR to filter and rank the semantic predications before expert manually review. We found that, although further development could improve the performance of the system, this filter provides a foundation to facilitate DSI discovery by avoid experts screen a larger reducing and enriching the set of semantic predications for expert review.

Acknowledgments

This research was supported by the University of Minnesota Informatics Institute On the Horizon Grant (RZ), Agency for Healthcare Research & Quality Grant (#R01HS022085-01) (GM), and University of Minnesota clinical and Translational Science Award (#8UL1TR000114-02) (Blazer). This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program, which is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

Reference

1. Peregoy JA, Clarke TC, Jones LI, Stussman BJ, Nahin RL. Regional variation in use of complementary health approaches by U.S. adults. NCHS data brief, no 146 Hyattsville, MD: National Center for Health Statistics. 2014.
2. National Center for Health Statistics About the National Health and Nutrition Examination Survey. Hyattsville (MD). 2009.
3. Dickinson A, Blatman J, El-Dash N, Franco JC. Consumer usage and reasons for using dietary supplements: report of a series of surveys. Journal of Americian College of Nutrition. 2014;33(2):176-82.
4. Kennedy DA, Seely D. Clinically based evidence of drug-herb interactions: a systematic review. Expert opinion on drug safety. 2010 Jan;9(1):79-124.
5. Kuhn MA. Herbal remedies: drug-herb interactions. Critical care nurse. 2002 Apr;22(2):22-8, 30, 2; quiz 4-5.
6. Gurley BJ, Swain A, Williams DK, Barone G, Battu SK. Gauging the clinical significance of P-glycoprotein-mediated herb-drug interactions: comparative effects of St. John's wort, Echinacea, clarithromycin, and rifampin on digoxin pharmacokinetics. Molecular nutrition & food research. 2008 Jul;52(7):772-9.
7. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in biology and medicine. 1986 Autumn;30(1):7-18.
8. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA Annu Symp Proc. 2006:349-53.
9. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003 Dec;36(6):462-77.
10. Kilicoglu H, Shin D, Fiszman M, Rosemblat G, Rindflesch TC. SemMedDB: a PubMed-scale repository of biomedical semantic predications. Bioinformatics. 2012 Dec 1;28(23):3158-60.
11. Zhang R, Cairelli MJ, Fiszman M, Rosemblat G, Kilicoglu H, Rindflesch TC, et al. Using semantic predications to uncover drug-drug interactions in clinical data. Journal of Biomedical Informatics. 2014;49:134-47.
12. Zhang R, Cairelli MJ, Fiszman M, Kilicoglu H, Rindflesch TC, Pakhomov SV, et al. Exploiting literature-derived knowledge and semantics to identify potential prostate cancer drugs. Cancer informatics. 2014.
13. Jonnalagadda SR, Del Fiol G, Medlin R, Weir C, Fiszman M, Mostafa J, et al. Automatically extracting sentences from Medline citations to support clinicians' information needs. Journal of the American Medical Informatics Association. 2013 Sep-Oct;20(5):995-1000.
14. Masseroli M, Kilicoglu H, Lang FM, Rindflesch TC. Argument-predicate distance as a filter for enhancing precision in extracting predications on the genetic etiology of disease. BMC bioinformatics. 2006 Jun 8;7:291.

Peer-Based Learning Workshops
==========

Instructors

Melissa Haendel, Nicole Washington, Chris Mungall

Title

The Monarch Initiative: Semantic phenotyping for translational medicine

Opening Summary

It is well-known that mutations in orthologous genes and genes in the same signaling pathway often manifest in similar phenotypes, and therefore study of variant phenotypes in model systems may provide insight into human gene function and understanding of disease. Traditionally, basic researchers and clinical geneticists have searched for and integrated this knowledge manually. However, with the rapid rise in output of genomic sequencing technologies and high-throughput phenotyping efforts, manual identification, integration, and evaluation of relevant phenotype data will quickly become intractable. Further, such data is difficult to interpret, requiring domain knowledge of each species' genetics, development, and even the specialized vocabulary used to describe anatomy, phenotype, and genotype. Furthermore, each organism is best suited for study of different biological phenomena, meaning an integrated view over all organisms is best for understanding relevance to human disease. Until recently, there has been a deficiency of tools that leverage cross-species phenotypes in translational research. The Monarch Initiative was created to provide an integrated data portal and suite of tools to address this need.

This workshop will cover the use of tools that leverage semantic phenotyping for the exploration of genotype-phenotype correlations for exome analysis, such as PhenIX and Exomiser, to aid disease-gene discovery. We will also focus the development and use of ontologies for capturing variant phenotypes and the evaluation of their quality using a phenotypic profile "sufficiency meter." Finally, we will highlight the visualization of human disease phenotypes in comparison to model systems to aid variant validation and model selection. We encourage participants to bring questions about the use of ontologies to capture phenotypes, VCF files together with phenotype annotations for analysis with Exomiser and PhenIX, and exploratory questions about their favorite animal model, disease, or gene.

Topic Outline

- Overview of (50%)
    - Semantic Phenotype Curation and Best Practices
    - Cross-species integrated ontologies: an introduction to Uberon/Uberpheno
    - Semantic analysis and visualization tools
    - Monarch Initiative portal and database as a resource
    - Exome analysis tools: Exomiser, PheniX
    - Ontology resources: TermGenie
    - Phenotype exchange standard and data sharing
- Small breakouts (50%)
    - Analyze participant or provided exome VCF and phenotype data together
    - How to curate phenotype data
    - Exploration and visualization of specific diseases and/or phenotypes within and across species

Length

- 90 minutes

Educational Objectives

Upon completion of this workshops, participants will be able to:

- Understand how phenotype ontologies are built and how their structure allows us to bridge across species and community-specific language to enable cross-species phenotype comparison
- Use Monarch initiative tools, including the phenotype similarity (Phenogrid) widget to interpret phenotypic similarity between genes, disease, and/or models
- Apply curation workflows to the recording of phenotypes
- Use the breadth of available phenotype data for comparison and analysis
- Integrate understanding of phenotype descriptions with variant prioritization during exome analysis
- Use exome analysis tools with phenotype data to facilitate interpretation of complex genotype/phenotype profiles

Who should attend

This workshop is recommended for anyone interested in learning more about the process of capture and computational analysis of phenotypes, and particularly researchers and clinical geneticists interested in learning how structured phenotypes can assist with variant prioritization in exome analysis. We also encourage policy-makers to attend to learn about how to make phenotypic research data interoperable and available for maximal reuse.

Prerequisites

- Familiarity with applications of one or more terminologies or ontologies (SNOMED-CT, GO, HPO)
- Basic understanding of exome analysis

Experience

- Project workshops, AMIA panels, academic and industry conference presentations, invited presentations, courses, and tutorials.

Use of the OMOP Common Data Model within the PCORI CDRN Initiative: Data Harmonization and Integration with PCORNet CDM

Michael E Matheny, MD, MS, MPH[1,2], Daniella Meeker, PhD[3], Michael G. Kahn, MD, PhD[4], Rimma Belenkaya, MA, MS[5], Lewis J. Frey, PhD[6], and Patrick Ryan, PhD[7]

[1] Geriatric Research Education and Clinical Care Center, TVHS Veterans Administration, Nashville, TN
[2] Departments of Biomedical Informatics, Medicine, and Statistics, Vanderbilt University Medical Center, Nashville, TN
[3] Department of Preventive Medicine, University of Southern California, Los Angeles, CA
[4] Department of Pediatrics, University of Colorado, Denver, CO
[5] Research Informatics Core, Albert Einstein College of Medicine, Yeshiva University, Bronx, NY
[6] Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC
[7] Janssen Research and Development, Raritan, NJ

Abstract

The PCORI Clinical Data Research Network initiative is in the process of integrating electronic health record and directly reported patient information from 11 participating networks of healthcare systems to pursue large, distributed observational cohort studies and pragmatic clinical trials. Each network has a unique starting point regarding design and implementation of their EHR and clinical data warehouse resources. A collaborative effort between OHDSI and 3 CDRNs (pScanner, PEDSnet, and NYC) with existing or in-progress deployments of the OMOP CDM have been developing crosswalks to the PCORNet CDM version 1 for OMOP version 4 and 5 in order to integrate with the emerging PCORNet CDM standard.

This workshop seeks to 1) provide a general overview of established common data models (I2B2, OMOP, PCORNet, Quality Data Model, and others) including relative strengths and weaknesses for categories of applications, 2) experience and challenges of transforming EHR data to the OMOP common data model, 3) ongoing work to provide cross-CDM interfaces between OMOP and PCORNet, and 4) development of ETL tools to help accelerate the transformation process for healthcare systems starting or in the process of data transformation, and 5) discussion of the process and ongoing efforts to update and support expanding use of a CDM from the curator's perspective (OHDSI)

In summary, this workshop is intended for audiences interested in using or implementing a secondary data use common data model within their institution, and serves as both an introduction to these types of models, and then proceeding into more detail with regards to the use of the PCORNet and OMOP CDMs for use within the CDRN network. Lessons learned and best practices will be discussed, tools being developed will be highlighted to help systems and groups perform these tasks more easily, and key challenges in application and re-use of CDMs will be described.

Topic Outline

- Overview of Secondary Data Use Common Data Models (Meeker)
    - I2B2
    - OMOP
    - PCORNet
    - Mini-Sentinel
    - Others
- OHDSI OMOP CDM Overview/History & user groups & tools available (Ryan)
- OMOP – PCORNet CDM Crosswalk – Development and Use (Belenkaya)
    - Rationale for Use of OMOP and then translation to PCORNet
    - Development of CDM Crosswalk – Challenges
    - Collaboration with OMOP-Based CDRN centers
    - Expectations for OMOP evolution
- OMOP & PCORNet Implementation Experience within National VA (Matheny)
    - Source Data Preparation
    - Controlled vocabulary coverage
    - Data Transformation
    - Computing Infrastructure (Incremental Loading)
    - Lessons Learned
- Tools to Support Data Transformation from an EHR or Clinical Data Warehouse to OMOP CDM (and other CDMs) (Kahn)
    - Open source data profiling tools
    - Open source ETL specifications writing tools
    - Open source data profiling tools
    - Use of social programming tools to align multi-institutional ETL rules

Workshop Length: 90 minutes, with each presentation 15 minutes, and 15 minutes for questions at the end

Educational Objectives

- Gain an understanding of the most commonly employed secondary use common data models, including a brief history of their evolution, strengths, and weaknesses
- Gain knowledge of resources necessary to transform EHR data to a common data model
- Understand the OMOP and PCORNet CDM
- Exposure to tools and applications to support transformation to OMOP CDM

Target Audience: Researchers, Healthcare System Leaders, IT Professionals

Content Level: 20% Basic, 60% Intermediate, 20% Advanced

Pre-Requisites: Experience with controlled vocabularies, interest in developing or using common data models to represent electronic health record data to conduct research

Instructor Experience:

Rimma Belenkaya: Ms. Belenkaya is a seasoned informaticist with years of experience and training in information modeling, knowledge engineering, application and data architecture, and statistical data analysis. She is currently the lead developer for the New York City CDRN.

Michael Kahn: Dr. Kahn is Professor of Pediatric Epidemiology in the Department of Medicine at the University of Colorado. He directs the Research Informatics group at Children's Hospital Colorado and the Compass Enterprise Data Warehouse group at the University of Colorado. He teaches database management systems at the College of Nursing and is a frequent guest lecturer both across the University of Colorado and at external Biomedical Informatics grand rounds and national webinars. His focus is on large-scale national data models and common data models. He has direct experience with HMORN VDW, i2b2, OMOP, Mini-Sentinel, and PCORnet CDMs. His research interests are in developing new models and methods for describing and measuring data quality within and across common data models.

Michael Matheny: Dr. Matheny is a practicing general internist and medical informatician, Associate Director of the TVHS VA Biomedical Informatics Fellowship, and Assistant Professor of Medicine, Bioinformatics, and Biostatistics at Vanderbilt University Medical Center. He has expertise in developing and adapting methods for post-marketing medical device surveillance, and has been involved in the development, evaluation, and validation of automated outcome surveillance statistical methods and computer applications. He leads the OMOP extract, transform, and load team within VINCI for the national VHA data, and is a Site PI for the pScanner CDRN led by Lucila Ohno-Machado. He is currently developing analytic tools using the OMOP CDM, and his other areas of research include natural language processing and health services research in acute kidney injury, diabetes, and device safety in interventional cardiology. He is currently supported by grant and contract funding from VA HSR&D, PCORI, FDA, and Astra Zeneca. He has a decade of experience guest lecturing in graduate informatics courses, and has participating in learning workshops at Academy Health.

Patrick Ryan: Dr. Ryan's research has focused on developing methodological strategies to study the appropriate use of observational healthcare data to identify and evaluate the effects of medical treatments. As a collaborator in Observational Health Data Sciences and Informatics (OHDSI) and investigator for the Observational Medical Outcomes Partnership (OMOP), I lead a collaborative research community focused on developing and evaluating analytical methods for active drug safety surveillance. Much of the methodological learning from this effort is directly transferrable to the proposed work. In my current position as the Head of Epidemiology Analytics at Janssen Research and Development, I am responsible for leading a team that conducts studies across an array of observational healthcare databases for multiple therapeutic areas. Throughout my research, I have demonstrated the ability to identify and develop novel solutions to maximize the use of observational healthcare data.

Lewis Frey: Dr. Frey develops novel algorithms and information systems for the purpose of discovery and data integration applicable to precision medicine. In addition to applying

novel machine learning to medically relevant data, his information systems approach combines the accumulated wealth of knowledge that exists in medical record systems with the vast amounts of molecular data being captured with high-throughput measurement technologies. Dr. Frey has developed and taught 4 informatics courses from 2007 to the current time, and has been a guest lecturer in other informatics graduate courses during the same time frame.

Title: The Use of Cytogenetic Data to Enable Drug Repurposing Studies

Authors: Zachary B. Abrams (B.S, Ohio State University, Columbus, OH) and Philip R.O. Payne (Ph.D., Ohio State University, Columbus, OH)

Background: Cytogenetic data in the form of karyotypes are commonly used in the diagnosis and treatment of many forms of cancer. Karyotype data are expressed in a text-based form that is not machine-readable. This limits the utility of these data for secondary use and research purposes. Utilizing the International System for Human Cytogenetic Nomenclature (ISCN), we developed a parsing and mapping system that allows karyotype data to be represented and analyzed in a computationally tractable manner. A Loss-Gain-Fusion model (LGF) was created that allowed us to represent each karyotype as a binary vector. Each cytogenetic region is represented three times (loss, gain, and fusion) in the model. We utilized the publicly available Mitelman database as a test-bed for analyses, focusing on problems related to drug repurposing[1].

Methods: Utilizing our computational model and the Mitelman database, we were able to successfully parse 98% of its karyotypes; of those parsed, 89.4% could be mapped into our binary Loss-Gain-Fusion model. We then classified karyotypes based on their disease labels and filtered out all diseases with less than 50 patients. We then selected genetic aberrations present in 20% or more of the population in which the cytogenetic event led to increased gene expression. Subsequently, we identified all genes in the affected region and found drugs that inhibited the function of the overexpressed gene using publicly available drug data in The Drug Gene Interaction Database (DGIdb). We performed a Pub Med literature search on these triplets selecting those in which diseases and drugs did not co-occur and where the disease and the gene co-occurred.

Results: We discovered 68,543 triplets containing (1) a disease, (2) an overexpressed gene, and (3) a drug that suppressed that specific gene. From this list, we discovered a total of 69 cancer disease-drug pairs that were not cited as co-occurring in the literature. Given this filtering process, were the drug and gene are related, the drug suppressed the gene, and the gene was implicated in the disease, it logically follows that the drug should be helpful in treating the disease.

Discussion: Our computational approach serves as a basis for new directions in drug repurposing leveraging existing and commonly available bio-molecular phenotype data. In order to validate our results, future laboratory-based testing will be conducted on a sub-set of our findings. The ability to link publicly available data sources is a central component of this work and emphasizes the importance of utilizing such data in conjunction with clinically-generated data sets so as to support in-silico hypothesis generation.

1. DvD: An R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. Clare Pacini1, Francesco Iorio1,2, Emanuel Gonçalves1, Murat Iskar3, Thomas Klabunde4, Peer Bork3 and Julio Saez-Rodriguez Bioinformatics (2013) 29 (1): 132-134.

DNA Methylation Data Analysis with SPIRIT-ML

Srisairam Achuthan[1], PhD, Kelsang Donyo, Zhuo Chen[2], Rama Natarajan[2] PhD, Joyce C. Niland[1], PhD, Ajay Shah[1], PhD

[1]Research Informatics Division, Department of Information Sciences, [2]Department of Diabetes and Metabolic Diseases Research, City of Hope, CA

Abstract

Epigenetic mechanisms have been implicated in many human diseases, including cancer and diabetes. DNA methylation is one type of epigenetic modification. Algorithms within SPIRIT-ML, a machine learning platform can be applied to group patient samples based on their DNA methylation profile with the goal of identifying differentially expressed genomic loci.

Introduction

Epigenetic mechanisms have been known to change the DNA and chromatin structure among unhealthy individuals relative to healthy controls. DNA methylation (DNAm), an epigenetic modification has been implicated in various types of cancer and diabetes. In mammalian cells, DNAm usually occurs in the context of CG dinucleotides (CpGs) and promoter DNAm is associated with gene repression. Significant effort and time are spent when individual data analysis methods are applied to analyze DNAm datasets generated by Infinium HumanMethylation450 BeadChip that are high dimensional in nature. SPIRIT-ML (Software Platform for Integrated Research Information and Transformation-Machine Learning) furnishes a streamlined approach to process the raw experimental data followed by clustering of patient samples based on their entire DNAm profile. The identified CpG sites of interest can help determine the genomic locations that are vital for identifying the functional significance of differentially expressed DNAm sites.

Methods

The SPIRIT-ML platform built on top of R, MATLAB and Pipeline Pilot provides clustering and classification algorithms for analyzing biomedical datasets such as DNAm. For validation purposes, the DNAm dataset from Charlotte Ling's lab was analyzed using SPIRIT-ML. The raw DNAm data provides the levels of methylated and unmethylated DNA at each CpG site across all patient samples. The raw data is converted to beta values which in turn can be converted to M-values defined as $\log_2$(beta/1-beta). Across all samples, the CpG sites are initially divided into at least three subsets – Class1, Class2 and Class3 based on their M-values being un/lowly methylated, moderately/highly methylated and lowly/moderately methylated; respectively. Clustering algorithms such as Cluster Clara and Cluster Agnes along with principal component analysis are then applied to these data subsets to discover the existence of any distinct clusters.

Results

Our goal is to identify the smallest number of CpGs within the DNAm profile that cluster the patient samples. We find that the principal components (first and second) divide certain data subsets (Class1 and Class2) into three distinct clusters. Based on only CpG sites in Class 1 we were not able to stratify patient samples based on before/after exercise. However, we find that a combination of moderately/highly methylated is able to stratify the patient samples based on pre-determined characteristics such as family history of type 2 diabetes.

Conclusions

SPIRIT ML is a functional machine learning platform that can discover and reveal patterns in datasets. We applied clustering algorithms to identify clusters of patient samples based on their DNAm profile. A key finding is that segregating the DNAm dataset into moderately/higly methylated sites can help stratify patient samples based on pre-determined characteristics compared to analyzing the DNAm profile as a whole. Most of the time, there is a need to identify smaller number of predictors (CpG sites) that can accurately predict the patient samples that are likely to fall into the observed number of clusters. The classification algorithms implemented within SPIRIT-ML can be utilized for this purpose. The predictors can help to determine functionally relevant genomic loci.

Automated Biospecimens Annotation Model for Scalable Mesothelioma Biobanking

Waqas Amin MD[1], Anil V. Parwani MD PhD[2], Jonathan Melamed MD[4], Gunasheil Mandava[1], Rahul Uppal[1], Carl Morrison MD[6], Carmelo Gaudioso MD PhD[6], Michael Feldman MD PhD[3], Harvey I, Pass MD[5], Rebecca Jacobson MD MS[1], Michael J. Becich MD PhD[1].

Department of Biomedical Informatics[1], University of Pittsburgh School of Medicine, Department of Pathology[2], University of Pittsburgh School of Medicine, Department of Pathology, University of Pennsylvania School of Medicine[3], Department of Pathology[4], New York University School of Medicine, Department of Cardiothoracic Surgery[5], New York University School of Medicine, Department of Pathology & Laboratory Medicine[6], Roswell Park Cancer Institute

Summary:
The National Mesothelioma Virtual Bank (NMVB) has provided a robust tumor-banking infrastructure for mesothelioma biospecimens to facilitate basic and clinical science discovery that will ultimately benefit the patients affected with this rare disease (1,2). NMVB has recently adopted a scalable, cost effective and automated biospecimens annotation approach using open source software solutions i2b2/SHRINE and TIES (Text Information Extraction System).

Background
Scientific advancements in the recent past have amplified the value of highly annotated specimens in the research of rare diseases. Biospecimens available to researchers that come with detailed clinicopathological annotations allow for more thorough and targeted research that will ultimately accelerate the coming era of personalized medicine. Currently, the NMVB is a leading resource of biospecimens (including tissue microarrays) to mesothelioma researchers. As it stands the specimens available through NMVB are well annotated and can be used by investigators for novel biomarker discovery, therapeutic intervention and outcomes measures research that will ultimately benefit mesothelioma patients.

Methods:
NMVB is now adopting a new scalable, cost effective and automated biospecimens annotation approach by adopting open source software solutions i2b2/SHRINE to develop a federated model tumor banking infrastructure (3,4). The model will be built upon standardized ontologies and data extracted and mined from electronic health records (EHR) and Cancer Registry (CR) at our partnering NMVB sites as an extension to our PCORI Clinical Data Research Network (CDRN) and NCATS Accrual to Clinical Trials (ACT) research programs. The data will be stored and queried via i2b2/SHRINE. We are employing TIES (Text Information Extraction System) to extract data from free text surgical pathology and radiology reports and augment the annotation process (5) and enrich it for biospecimen dependent translational research efforts.

Results:
By implementing i2b2 software implemented at each of the collaborative sites, SHRINE will connect the sites to create a peer-to-peer, federated network for NMVB. This endeavor will be unique in the fact that the network being developed will also have the capacity to facilitate tissue banking alongside permissioned data sharing. The new architecture will employ TIES, EHR, and CR to populate the i2b2 instances present at each collaborative site. Through the ETL process, important information will be stored in an i2b2 data warehouse (at each site), that will then be made accessible through SHRINE to streamlined cohort identification and tissue and associated clinical data request fulfillment for researchers. TIES and CR already have their own methods of extracting structured data from unstructured portions of pathology reports and EHR records and these will be utilized in this new model as well.

Discussion:
This new NMVB federated biospecimen annotation model will allow for permissioned data sharing that will be more cost effective than manual annotation. A subsequent study will analyze the cost savings and the data quality improvements that this new model provides. By maximizing the utility and efficiency of a federated data and tissue network, we also aim to enable sharing of biospecimens for nova biomarker and therapeutic discovery that will ultimately benefit the patient population.

References:

1. Amin W, Parwani AV, Melamed J, Raja F, Pennathur A, Valdevieso F, Whelan NB, Landreneau R, Luketich J, Feldman M, Pass HI, Becich MJ. National Mesothelioma Virtual Bank: A Platform for Collaborative Research and Mesothelioma Biobanking Resource to Support Translational Research, Lung Cancer International, vol. 2013, Article ID 765748, 9 pages, 2013. doi:10.1155/2013/765748.

2. Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, Schmandt L, Winters SB, Milliken E, Kim P, Whelan NB, Farhat G, Melamed J, Taioli E, Dhir R, Pass HI, and Becich MJ. The development and deployment of Common Data Elements for tissue banks for translational research in cancer - an emerging standard based approach for the Mesothelioma Virtual Tissue Bank. BMC Cancer. 2008 Apr 8;8:91. PMID: 18397527 PMCID: PMC2329649

3. Anderson N, Abend A, Mandel A, Geraghty E, Gabriel D, Wynden R, et al. Implementation of a deidentified federated data network for population-based cohort discovery. Journal of the American Medical Informatics Association : JAMIA. 2012 Jun;19(e1):e60-7. PubMed PMID: 21873473. Pubmed Central PMCID: 3392860.
4. McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, et al. SHRINE: enabling nationally scalable multi-site disease studies. PloS one. 2013;8(3):e55811. PubMed PMID: 23533569. Pubmed Central PMCID: 3591385.

5. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M. caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. Journal of the American Medical Informatics Association : JAMIA. 2010 May-Jun;17(3):253-64. PubMed PMID: 20442142. Pubmed Central PMCID: 2995710.

Assessing the plausibility of drug interactions signals learned from the EHR

Juan M. Banda, PhD[1], Howard Strasberg, MD, MS[2], Nigam H. Shah, MBBS, PhD[1]
[1]Stanford Center for Biomedical Informatics Research, Stanford, CA; [2]Wolters Kluwer Health, Philadelphia, PA

Abstract

Early identification of drug-drug interactions (DDI) is of vital importance because over 30% of all adverse drug reactions are due to drug interaction, resulting in significant morbidity every year. In previous work we presented an approach for identifying DDI signals directly from EHRs using adjusted disproportionality ratio thresholds, resulting in the first published database of adverse event rates among patients on drug combinations based on an EHR corpus. Although it is feasible to identify DDI signals and to estimate the rate of adverse events directly from clinical text, the positive predictive value of the interactions in the database is unknown and difficult to estimate. In order to address this problem, we define an alternative conservative scoring criterion to reduce the false positive rate of drug combinations classified as interactions. We then compare the resulting set of predictions with the original ones, and present three metrics to assess the quality of the predictions.

Introduction

In our work we utilize the database of population event rates of DDIs introduced by Iyer et. al[1] and propose a more restrictive constraint on their proposed method with the objective of reducing the false positive rate of DDI signals identified. The restrictive constraint introduced assumes that at least one of the drugs is always present, in contrast to the original method that had an at most constraint for the bottom relationships of our 2 x 2 contingency tables. We also propose three metrics that evaluate the predicted signals against other well-known and validated DDI prediction methodologies to assess their plausibility and be able to score each of the proposed methods.

Methods

First we quantify our modified method by providing an area under the receiver operator characteristic (ROC) curve and comparing with previous results. We then develop three metrics that evaluate the predicted DDI signals in terms of: 1) Validation of predictions against new DDI reports added to public sources such as Drugbank, FAERS, and Drugs.com after the predictions were made, 2) Comparison against other predictive methods that do not rely on any EHR data such as: TWOSIDES (Tatonetti et al.[2]) and INDI (Gottlieb et al.[3]), and 3) Automated Plausibility Evaluation using drug targets with adapted method from EU-ADR project[4]. These metrics are developed with the purpose of assessing the feasibility of the predictions made by the original method and the conservative one we propose. Figure 1 provides an overview of our methods.

Results

We find that the conservative method finds 2,770 potential DDI signals versus 5,983 potential DDI signals found with the original method. In an evaluation using a held out test set the new method has an AUROC of 71.9% as compared to the AUROC of 82.3% for the original approach. While the AUROC value for our proposed method is lower, we are working under the hypothesis that with the method being more conservative, the analysis we perform will show that its predictions are proportionally more plausible. We have an overlap of 521 DDIs as being identified by both methods. We evaluated the non-overlapping predictions found in method 1 and method 2, as well as their overlap independently. The reported DDI signals found in the database of Iyer et al.[1] are generated with data up to 2011, so for the first assessment metric we have evaluate the reports added on FAERS, Drugbank, drugs.com dated after 2012. A similar evaluation is performed with the predictions made by TWOSIDES[2] and INDI[3]. Our preliminary results are in Figure 2.

Discussion

It is not possible to test all predicted interactions between drugs in an experimental setting. However, a DDI signal prediction method can be effectively assessed by its ability to correctly predict DDI signals that have subsequently appeared in FAERS, Drug Bank and Drugs.com reports in the last two years. Overlap with other non-EHR prediction methods, and plausibility evaluation using drug targets are other ways of assessing the correctness of a predicted DDI. We note that when assessing the sets of predicted DDI signals through the presented metrics, we can also examine the consensus between the different metrics for a given DDI and prioritize experimental validation.

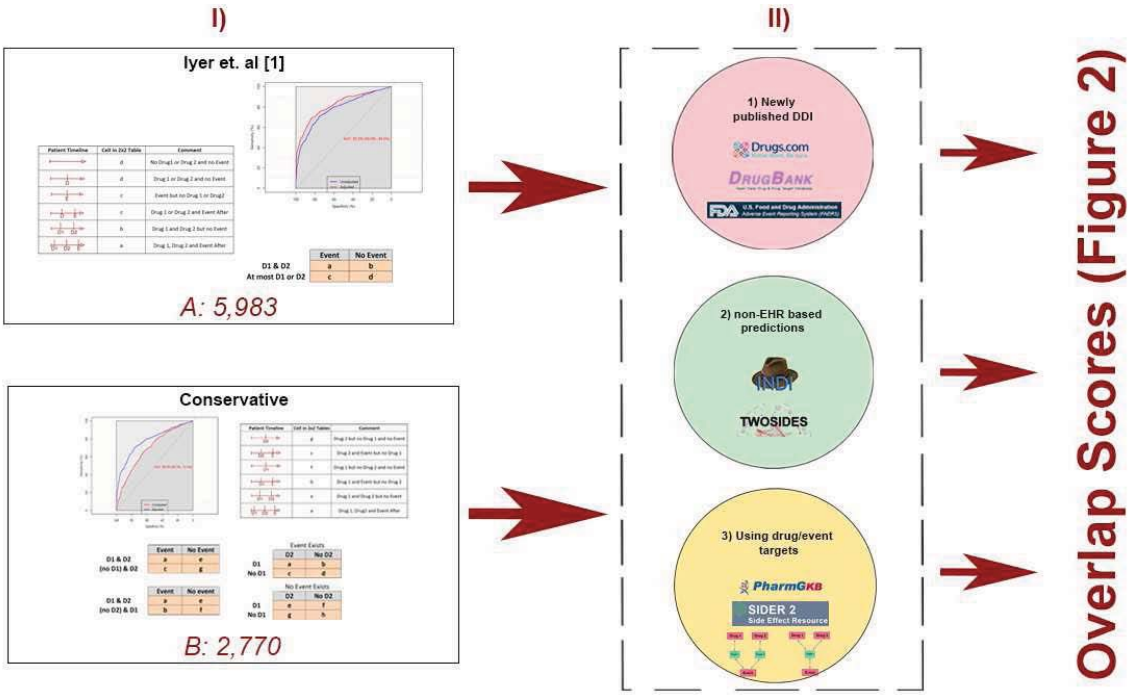


Figure 1. Section I) indicates the filtering methods for the DDIs found in the database of Iyer et al[1] .Section II) shows the three separate evaluation metrics for which the results are shown in figure 2.

	1			2		3
	Drugbank*	**Drugs.com***	**FAERS**	**INDI**	**TWOSIDES**	**Targets**
A - B	113	162	492	457	514	*In progress*
A ∩ B	9	5	60	41	158	*In progress*
B - A	41	27	207	181	702	*In progress*

Figure 2. Label A indicates the method presented on Iyer et al.[1], label B indicates our restrictive method. The number in each cell represents the total number of DDIs that overlap with each source/method. Section 3/Targets results are a work in progress and, if selected, they will be ready in time.

References

1. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc. 2014;21(2):353-362.
2. Tatonetti, NP, Ye, PP, Daneshjou, R, Altman, RB. Data-driven prediction of drug effects and interactions. Sci. Transl. Med. 4, 125ra31 (2012).
3. Gottlieb A, Stein GY, Oron Y, et al. INDI: a computational framework for inferring drug interactions and their associated recommendations. Mol Syst Biol 2012;8:592.
4. Bauer-Mehren, A. et al. Automatic filtering and substantiation of drug safety signals. PLoS Comput. Biol. 8, e1002457 (2012)

Identification of zoonotic disease clusters by integrating phylogeography

Rachel Beard, Matthew Scotch PhD, MPH
Arizona State University, Tempe, AZ, USA

Abstract
We describe a novel geospatial cluster algorithm for identifying centers of viral outbreak and spatial diffusion using parameters from phylogeography. Phylogeography uses genetic sequences in addition to temporal and geographical metadata regarding viral host and collection date to analyze viral diffusion. Here, we reconstruct transition rates among discrete locations using sequences to introduce a new weighting scheme for Local Moran's I cluster statistic. We implement our approach using West Nile virus.

Introduction
Over the last few decades, there has been an emergence or re-emergence of zoonotic viruses such as novel influenza viruses, sudden acute respiratory syndrome (SARS), and Ebola virus. This presents a challenge for public health agencies to quickly identify and respond to emerging threats in order to curb outbreaks and reduce morbidity and mortality.

Background
Clustering has been used previously to identify high risk areas, using algorithms such as Kuldorf's Spatial scan statistic or Local Moran's I.[1, 2] However, these models do not account for complex interactions among evolutionary, epidemiological and ecological factors that influence the spread of viruses, such as movement of reservoir hosts, population density, or the local environment. [3-5]

We describe a novel geospatial cluster algorithm for virus outbreak detection using parameters from phylogeography. Here, we reconstruct transition rates among discrete locations using virus sequence data to introduce a weight-based extension of the frequently used Moran's Local I cluster statistic. This work is important to build towards the implementation of predictive models to anticipate risk areas for virus outbreaks. We describe and implement our approach using West Nile virus (WNV) data in Texas.

Methods
We obtained WNV from ArboNet,[6] a Federal surveillance system which documents reported cases of WNV virus in humans, birds, mosquitoes and other mammals. We aggregated all reported cases by county over a span of ten years from 2003-2012. We extracted virus sequences from Genbank[7] using the taxonomy search tool (where concept name = West Nile virus), and specifying complete genome sequences collected from Texas during 2003-2012. Additionally, one criteria for sequence selection was geographic meta-data at the county level in the GenBank record. Using BEAST v. 1.8,[8] we performed a reversible ancestral reconstruction based on 22 discrete (county) locations from 63 sequences to create a transmission rate matrix. Using ArcGIS 10.1 we performed clustering analysis using binary weights (1 or 0 if counties within Texas are directly adjacent or not, respectively) and compared it to a weighting scheme based on the transition rates among discrete locations. As the number of counties with available viral sequences did not match the number of affected counties, we populated any unsampled counties with the transition rate of its nearest sampled county. We standardized the data to accommodate for underreported case data. We also prepared a visualization based on a subset of the extracted sequences collected from avian species to highlight patterns of diffusion throughout Texas, as birds have been indicated in the spread and reintroduction of WNV within regions.[9]

Results
All counties identified with highly significant clustering (p value<0.05) using viral transition rates overlapped with those identified using a binary weight, though were restricted geographically to the central-western portion of the state (Figure 1, A and B). This aligns with previous phylogeographic work performed by Pybus et al.,[10] in addition to our own analysis using sequences collected from avian species, which indicates the central-western portion of Texas as seeding ground for several nearby regions of the US.

Discussion
Using this modified clustering method, researchers could potentially identify areas with high rates of viral spread with greater specificity. Lending additional support to the importance of the clustering occurring in central western counties, previous studies have found that these areas have an elevated relative risk in the human population.[11] For future work we will address potential predictors of clustered areas, using techniques such as spatial regression of

attributes associated with WNV cases, and predictive data mining. Finally, we will expand on this work by considering multiple viruses and geographic areas and computationally address limitations to sparse sampling data.

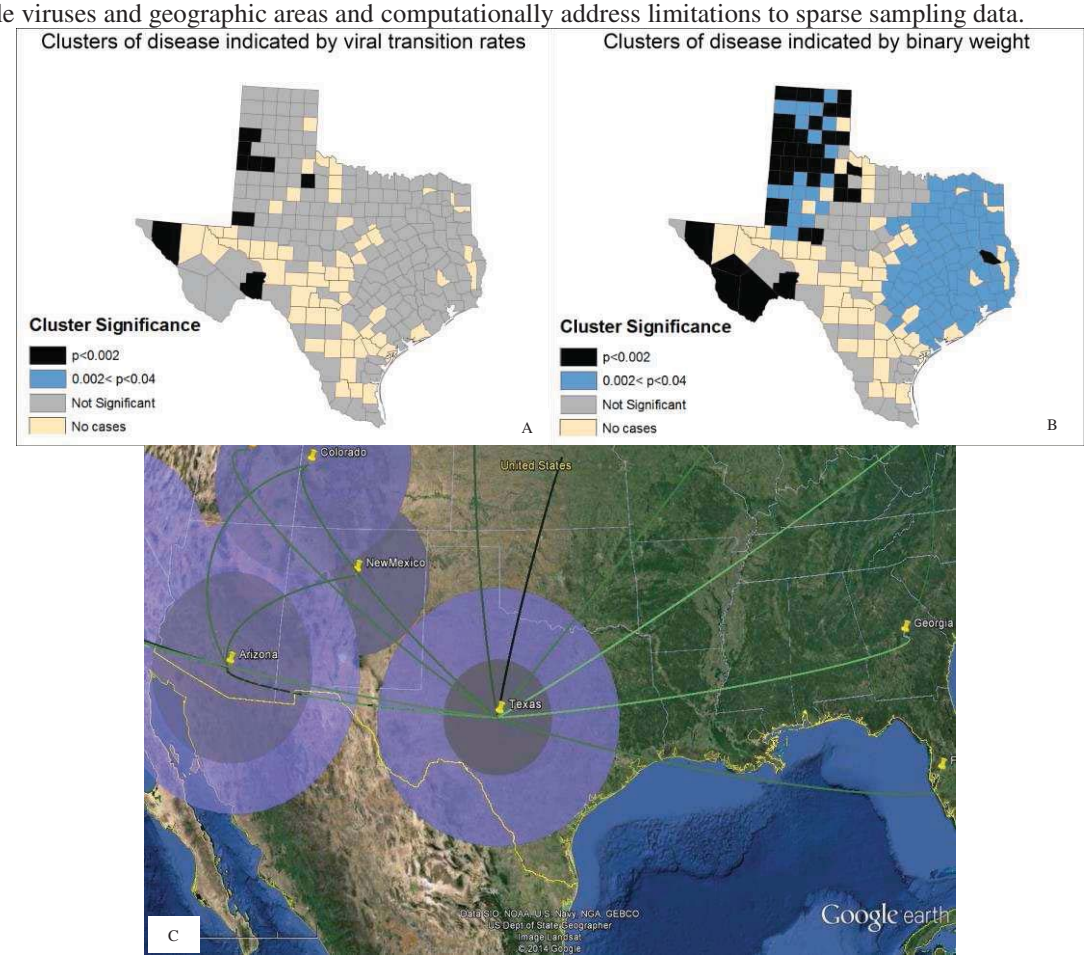


Figure 1. A) Clusters identified using viral transition rates. B) Clusters identified using the binary weights generally used in Local Moran's I. C) The map depicts the spreading pattern of West Nile virus inferred from sequence samples from avian species during the study period. Here, the spread of West Nile virus is shown after arrival in Texas in 2003, and the progression to other nearby regions over the next several years. The rings around the central locations from which sequences were sampled indicate a large number of ancestral states for a span of time.

References

1. Song C, Kulldorff M. Power evaluation of disease clustering tests. International journal of health geographics. 2003;2(1):9-.
2. Sugumaran R, Larson SR, Degroote JP. Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. International journal of health geographics. 2009;8(1):43-.
3. Morse SS, Mazet JAK, Woolhouse M, Parrish CR, Carroll D, Karesh WB, et al. Prediction and prevention of the next pandemic zoonosis. The Lancet. 2012;380(9857):1956-65.
4. Bordier M, Roger F. Zoonoses in South-East Asia: a regional burden, a global threat. Animal health research reviews / Conference of Research Workers in Animal Diseases. 2013;14(1):40-67.
5. Wiwanitkit V, Shi B, Xia S, Yang GJ, Zhou XN, Liu J. Research priorities in modeling the transmission risks of H7N9 bird flu. Infectious diseases of poverty. 2013;2(1):17.
6. <Arbonet.pdf>.
7. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic acids research. 2012;40(Database issue):D48-53.
8. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular biology and evolution. 2012;29(8):1969-73.
9. Ozdenerol E, Taff GN, Akkus C. Exploring the spatio-temporal dynamics of reservoir hosts, vectors, and human hosts of West Nile virus: a review of the recent literature. International journal of environmental research and public health. 2013;10(11):5399.
10. Pybus OG, Busch MP, Delwart EL, Suchard MA, Lemey P, Bernardin FJ, et al. Unifying the spatial epidemiology and molecular evolution of emerging epidemics. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(37):15066-71.
11. Murray KO, Ruktanonchai D, Hesalroad D, Fonken E, Nolan MS. West Nile virus, Texas, USA, 2012. Emerging infectious diseases. 2013;19(11):1836.

Quantifying Geo-imputation error: Using Gaussian Geostatistical Simulation (GGS) to Dis-aggregate Zip Code Data & Estimate Positional Error.

Jess J. Behrens, M.Sc.[1], Adam R. Pah, Ph.D.[1], Abel N. Kho MD, MS[1]

[1]Northwestern University, Feinberg School of Medicine

Abstract: Despite privacy concerns, an increasing amount of research is directed at identifying the role that environment plays in the disease process. Using the 2008 North Carolina Voter dataset (~6 million records) we geocoded addresses, aggregated by zip code, and then disaggregated to block group using Monte Carlo & Gaussian Geostatistical Simulation (GGS) techniques. We identified a high degree of positional accuracy when comparing expected to observed address locations.

Introduction & Background: There is a growing body of literature examining the role that geography plays in the disease process (Diez Roux et. al., 1995; Morland et. al., 2002). Given the HIPAA regulations covering patient privacy, virtually all geographic analyses of patient diagnoses are limited to aggregation by coarse geographic regions such as county or 3 digit zip code. However, several studies have demonstrated that geographic analysis by zip code is inexact at best and misleading at worst, given that there is often greater variability in socio-economic and demographic data within each zip code than exists between zip codes in a given region (Krieger et. al. 2002). As such, there has been significant effort toward developing statistical methods for disaggregating patient records to smaller, more homogenous, geographically coincident census based units, such as census tract or block group (Mennis, 2003; Poulsen & Kennedy, 2004; Henry & Boscoe, 2008). All of these methods rely heavily on assumptions about the relationship between patient demographics and the demographics of the associated census geographies, but few have ever tested either the accuracy of those assumptions or if factors such as the effect of moving between different levels of geographic scale (zip code vs. census tract vs. census block group) and population density via urban/rural classification play a significant role in case assignment (Henry & Boscoe, 2008).

Perhaps the main reason for this oversight is due to the absence of a cohesive method for estimating case assignment error that is geographically continuous. All error estimates up to now evaluate the assignment of a case to a geographical unit, which is a discrete measurement. Gaussian Geo-statistical Simulation methodology provides an opportunity to examine case assignment error using a geographically continuous surface (Dietrich & Newsam, 1993). As such, we set out to examine this novel method for disaggregating patient location from zip code to coincident census geographic units.

Methods: ArcGIS 10 was used to address match the publicly available North Carolina voter data set, comprising ~6 million records, and to select a subset of 64 zip codes including both urban and rural areas in and around Winston-Salem. Voter addresses were aggregated by zip code and three demographic features of each voter was recorded: gender, age, and ethnicity. Geographically coincident block groups were assigned to each zip code by block group centroid and single origin ethnicity, age group, and gender totals were recorded for each. North Carolina voter totals aggregated by study area zip code were then assigned to the associated block groups using a monte carlo simulation that weighted voter block group assignment using underlying block group demographic counts. Each monte carlo ensemble included 10 separate simulations, each with 1000 iterations, using 1, 2, 10, 15, 20, 25, 30, 40, & 50% of the total voters in the study area. Voters were randomly selected from the data set for each level grouping of voters by percent. Commensurate with standard GGS methodology, the average number of voters assigned to each block group was kriged using ArcGIS 10 geo-statistical analyst. A separate krig was optimized for each ensemble of 10 monte carlo simulations based on the percent of voters included in that subset, 1 – 50%. The kriged surface was then combined with the average number of voters assigned to each block group and along with the error term produced by the krig in a Gaussian Geo-statistical Simulation. GGS uses the semi-variogram developed during the kriging process along with the average number of voters assigned to each block group during the monte carlo simulation and the error term from the krig to simulate and interpolate a continuous, raster based estimate of average/standard deviation voter values for the area between the block group centroids.

Results and Discussion: Examination of the results demonstrated a high degree of accuracy in voter location assignment, whether evaluated using the z-score in aggregate by census block group or address. While some z-scores were $p < 0.01$, the vast majority fell within the normal range, indicating that it is possible to impute location from aggregate de-identified data using aggregate census demographic values as surrogates for probable location. A marked shift in accuracy was uncovered as the model moved from rural (low density) to urban (high density) areas, however, and will need to be taken into consideration when applying the method.

Bibliography

Barr, R.G., Diez Roux, A.V., Knirsch, C.A., & Pablos-Mendez, A. 2001. "Neighborhood Poverty and the Resurgence of Tuberculosis in New York City, 1984 to 1992," American Journal of Public Health, 91 (11): 1783-1789.

Carretta, H.J. & Mick, S.S. 2003. "Geocoding public health data," American Journal of Public Health, 93 (5): 699-700.

Dietrich, C.R. & Newsam, G.N. 1993. "A fast and exact method for multidimensional Gaussian stochastic simulations," Water Resources Research, 29 (8): 2861-2869.

Diez Roux, A.V., Nieto, F.J., Tyroler, H.A., Crum, L., & Szklo, M. 1995. "Social inequalities and atherosclerosis: the ARIC Study," American Journal of Epidemiology, 144: 1048-1057.

Diez Roux, A.V. 1998. "Bringing context back into epidemiology: variables and fallacies in multilevel analysis," American Journal of Public Health, 88: 216-222.

Goovaerts, P. 1997. Geostatistics for Natural Resources Evaluation, Oxford University Press, 483 pp.

Henry, K. & Boscoe, F. 2008. "Estimating the accuracy of geographical imputation," International Journal of Health Geographics, 7 (3): online.

Isaaks, E.H. & Srivastava, R.M. 1989. An Introduction to Applied Geostatistics, Oxford University Press, 561 pp.

Krieger, N., Chen, J.T., Waterman, P.D., Soobader, M.J., Subramanian, S.V., & Carson, R. 2002. "Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project." American Journal of Epidemiology, 156: 471-482.

Mennis, J. 2003. "Generating Surface Models of Population Using Dasymetric Mapping," The Professional Geographer, 55 (1): 31-42.

Morland, K., Wing, S., Diez Roux, A.V., & Poole, C. 2002. "Neighborhood characteristics associated with the location of food stores and food service places," American Journal of Preventive Medicine, 22 (1): 23-29.

Poulsen, E. & Kennedy, L. 2004. "Using Dasymetric Mapping for Spatially Aggregated Crime Data," Journal of Quantitative Criminology, 20 (3): 243-262.

Ruston, G., Armstrong, M.P., Gittler, J., Greene, B.R., Pavlik, C.E., West, M.M., Zimmerman, D.L. 2006. "Geocoding in cancer research: a review," American Journal of Preventive Medicine, 30: S16-24.

Schootman, M., Sterling, D., Struthers, J., Yan, Y., Laboure, T., Emo, B., & Higgs, G. 2007. "Positional Accuracy and Geographic Bias of Four Methods of Geocoding in Epidemiologic Reasearch," Annals of Epidemiology, 17 (6): 464-470.

Inter-Network Cluster Replication: A Case Study in Co-Occurring Comorbidities

Suresh K. Bhavnani, PhD[1], Bryant Dang, BS[1], Shyam Visweswaran, MD, PhD[2], Rohit Divekar, MBBS, PhD[3]
[1]Inst. for Translational Sciences, Inst. for Human Infections and Immunity, Univ. of Texas Medical Branch, Galveston, TX; [2]Department of Biomedical Informatics, Univ. of Pittsburgh, Pittsburgh, PA; [3]Division of Allergic Diseases, Mayo Clinic, Rochester, MN

Abstract

Although networks have been extensively used to identify complex associations in biomedical data, few studies have replicated those associations by comparing them across independent datasets. Here we describe a method for comparing the degree of cluster similarity between two networks that represent the co-morbidities of hip-fracture (HFx) patients in 2010, and in 2009, retrieved from the Medicare database. We demonstrate a significance test for this similarity measure, which can be used to test the replicability of clusters across networks.

Introduction

While networks have been effective in identifying complex associations among subjects and variables[1], few of these studies have replicated their results in another dataset. As little is known about how co-morbidities co-occur and replicate in 30-day readmitted HFx patients, our research question was: *"How to measure the degree and significance of comorbidity cluster similarity between two cohorts of HFx patients in the Medicare database?"*

Method

We extracted all 30-day readmitted HFx patients without joint replacement from the 2010 and 2009 Medicare database, in addition to 69 of their comorbidities determined to be critical[2] in the elderly. Using a cumulative frequency distribution, we excluded 32 comorbidities that together accounted for <1% of the patients. Next, we used bipartite networks (consisting of patients and comorbidities) and co-cluster modularity[1] (representing degree of clusteredness) to analyze how the remaining 37 comorbidities clustered in 2010, and in 2009. Finally, we measured the cluster similarity across the two years by (1) calculating the *degree of inter-network cluster similarity* based on the formula: number of co-occurring pairs of comorbidities in a cluster in both years / the number of co-occurring pairs of comorbidities in a cluster in any year (where 0=no inter-network cluster similarity, and 1=complete inter-network cluster similarity), and (2) testing its significance by comparing it to a distribution of the same measure generated from 1000 random permutations of the 2010 and 2009 networks.

Results and Conclusion

Co-cluster modularity was similarly high in both years (2010=0.39, 2009=0.41), and the *degree of inter-network cluster similarity* between the two years was highly significant ($p<0.001$, readmission=0.33, random mean=0.063). Furthermore, as shown in Figure 1, an inspection of the two network layouts revealed that although they differed in the total number of clusters (2010=5, 2009=7), the most prevalent comorbidities (accounting for majority of the patients in their clusters and labeled in the networks) were identical in 4 clusters in both years. These replicated patterns of prevalence and co-occurrence pinpoint comorbidities that could in future research help to determine follow-up care with the goal of reducing 30-day readmission in HFx patients. The results suggest that the above approach should be effective for testing the replicability of clusters in biomedical data ranging from molecules to comorbidities.

References

1. Newman M. Networks: An Introduction. Oxford University Press; 2010.
2. Pope GC, Kautter J, Ellis RP, et al. Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model. Health Care Financ Rev. 2004 Summer;25(4):119-41.

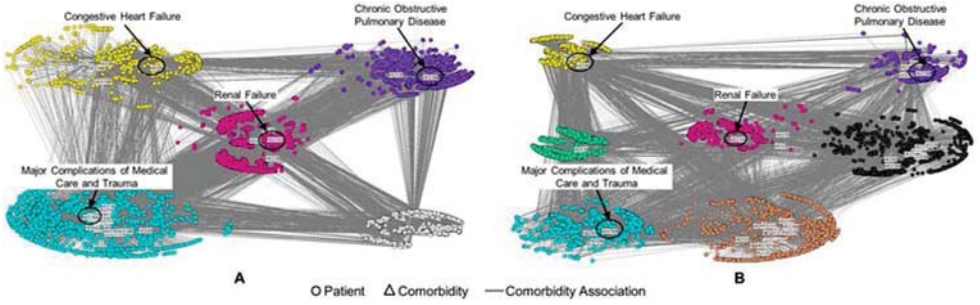


Figure 1. Bipartite networks showing clusters in 2010 (**A**), and in 2009 (**B**) using an exploded 2D layout view which pulls apart clusters identified through modularity, while preserving the size and spread of the nodes in each cluster.

A data safe haven to securely bring analysis to distributed cancer genome data

Francisco M. De La Vega[1], Ying Wu[1], Tal Shmaya[1], Thomas Schlumpberger[1], James Wiley[2], Akshay Patel[2], and Raja Hayek[2]
Annai Systems, Inc. [1]Burlingame, CA, and [2]Carlsbad, CA

To target and personalize cancer therapies to the genomic aberrations present in a particular patient's tumor, researchers need a comprehensive catalogue of the somatic mutations that arise during the formation of malignant tumors, and models of how these alterations interact to give rise to tumor phenotype. This requires analysis of genome data from large samples of patients' genomes to identify driver mutations in the "tail end" of the frequency distribution. Community genomics data sets from the TCGA and ICGC projects represent a valuable resource to which researchers can add their own data to gain statistical power in their analyzes. The current issue to this methodology is the highly fragmented storage of public and private data and the inefficient access to public data. Researchers spend weeks to months downloading hundreds of terabytes of data from central repositories before computations can begin. What is needed is a data "safe haven" where researchers can bring compute to the reference data without the need to incur in bulky data transfers or duplicative storage costs, in an environment that protects the privacy of the patients' data. Annai-ShareSeq is a genomic data safe haven that provides a technical solution for storing, handling and analyzing identifiable genomic data. This resource leverages Annai-GNOS, the technology we developed to create and manage the CGHub TCGA repository together with UCSC, and that is being used in the ICGC Pan Cancer Analysis of Whole Genomes project, and combines it with a high-performance compute environment and an array of tools to process and analyze genomic data. Built using a walled garden approach, where the data is stored, processed and managed within the security of the system, ShareSeq avoids the complexity of assured end point encryption. GeneTorrent, our fast and secure file transfer mechanism, enables researchers' private information to be transferred into the walled garden simply and securely to combine it with the public datasets. ShareSeq differs dramatically from the traditional cloud in two features: (i) formal mechanisms and a service level agreement to store protected identifiable genomic data securely and safely, built into the system from the ground up; (ii) the system is specifically designed for genomic computing over large shared data sets supporting common bioinformatics workflow tools; (iii) Fast download and access to raw genomic information and its metadata through a webservices interface; (iv) access controls leveraging federated authentication systems that Data Access Committees utilize to authorize access to the restricted data; and (iv) ability to plug in provenance management tools such as Synapse to enhance analysis reproducibility. ShareSeq is initially hosting raw, normalized, and processed data from the International Cancer Genome Consortium (whole genome sequence, exomic, and transcriptomic data). We envision that over time ShareSeq will host an increasing number of high value reference genomic public datasets and add standards-based interfaces promoted by the Global Alliance of Genomes and Health to allow broader data discovery and sharing.

Development of Standardized Common Data Elements for Biobanking of Human Biospecimens

Helena J. Ellis[1], Mary-Beth M. Joshi[2], Aenoch J. Lynn[3], MPH Anita C. Walden[4]

[1]Duke Biobank, Duke University, Durham, NC; [2]Department of Surgery, Duke University, Durham, NC; [3]Buck Institute for Research on Aging, Novato, CA; [4]Duke Translational Medicine Institute, Duke University, Durham, NC

Abstract

Biobanking is a critical tactical component of clinical and translational research. However, the field of biobanking lacks standardization in practice as well as in informatics. Employing a consensus driven process, the Duke Biobank developed a biobanking standard terminology for use in their enterprise-wide biospecimen information management system and plans to move the terminology forward on a national front.

Introduction

According to Carolyn Compton, former Director of the National Cancer Institute (NCI) Office of Biorepositories and Biospecimen Research, "biospecimens and pathologists are the center of the translational medicine universe". In order to conduct translational research and make headway into Personalized or Precision Medicine, researchers must have clinical information on the participant, as well as the conditions experienced by the biological sample during its lifecycle that may affect downstream processing. In 2014, the NCI identified the lack of standardized, high quality biospecimens with essential clinical data as a significant roadblock to cancer research. Furthermore, rarely does a single biobank have sufficient samples of a specific disease to achieve required statistical power, thereby requiring merging samples and data from different biobanks and even different organizations. Biospecimen data and metadata must be collected and stored, and be readily retrievable in a standardized format. Otherwise, the simplest query to join clinical data with sample data becomes, at best, a tedious effort with loss of time and information. At worst, linkage becomes an impossible task.

Background

Duke, like most academic medical centers, has dozens if not hundreds of diverse biobanks with informatics systems that have proliferated and developed in silos and without regard to semantic interoperability. As part of an initiative to implement an enterprise-wide biobanking informatics system, the Duke Biobank developed a biobanking standard terminology.

Methods

The scope of the project was defined as the lifecycle of the biospecimen, including participant consent, sample collection, processing and storage, through distribution and testing. Employing a consensus driven process, working groups were established for each of five domain areas: Sample Collection and Storage, Tracking and Non-Chemical Processing, Chemical Processing and Derivatives, Complex/Omics Data and Clinical Data. Each working group included an Informaticist and a Facilitator in order to tightly manage scope and communication between groups. Beginning with authoritative sources, and existing Duke data elements, each working group identified and defined terms within their scope. An Oversight Committee made final decisions and settled disputes. The work product from all working groups underwent an online Duke-wide public comment period.

Results and Discussion

Over a nine month period, over 500 data elements and definitions were identified and developed. The data elements have been further refined and revised in order to harmonize with the out-of-the box terminology in the purchased commercial software. Duke is working with the International Society of Environmental Repositories (ISBER), the College of American Pathologists (CAP), as well as commercial and academic biobanks to submit these Biobanking Common Data Elements to a national standards organization, such as LOINC, for management, curation and distribution.

The use of a single terminology will promote interoperability and reduce ambiguity and errors, while encouraging biobanking best practices in regards to tracking of key pre-analytical variables. The terminology described here may be useful to other institutions facing structuring and standardization of biobanking practices and technologies.

Extension of RxNorm for Medication Clinical Decision Support

Robert R. Freimuth, PhD[1,2], Kelly Wix[3], PharmD, RPh, Mark Siska, RPh[3], Christopher G. Chute, MD, DrPH[1]

[1]Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, [2]Office of Information and Knowledge Management, [3]Pharmacy Services Information Systems; Mayo Clinic, Rochester, MN

Summary

Previously, we observed a high level of semantic pre-coordination in terms representing dose form and route of administration within RxNorm. We propose extensions to RxNorm, including 9 new properties and associated value sets, which would significantly improve the ability to create and maintain drug lists for medication CDS.

Introduction and Background

Standardized representations of dose form and route of administration are needed to facilitate management of medication lists for clinical decision support (CDS) rules. Previously, we observed a high level of semantic pre-coordination in terms representing these concepts from RxNorm, RxTerms, and NDF-RT, which complicated the use of these standard resources for medication CDS[1]. We sought to address these limitations by creating a set of properties and associated value sets that could represent the concepts observed in the source terminologies through post-coordination. These proposed extensions will improve the ability to use RxNorm for medication CDS.

Methods

The RxNorm and RxTerms data files were downloaded from the NLM Unified Medical Language System (UMLS) web site, and concepts related to dose form were extracted. The NDF-RT "Dose Forms" hierarchy (NUI N0000010010, NDFRT_KIND property of "DOSE_FORM_KIND") was downloaded via the NLM NDF-RT REST API. Each term was reviewed and semantically decomposed into post-coordinated concepts, as needed. The resulting concepts were grouped into categories that extended the existing properties and relationships in RxNorm and NDF-RT. The terms listed in the U.S. Pharmacopeial Forum (2009) and those included in the FDA standard for dosage form (OID 2.16.840.1.113883.3.26.1.1.2) were reviewed and used to extend the concepts that were derived from RxNorm, RxTerms, and NDF-RT. The new value sets were mapped to the original terms.

Results

Most of the terms from RxNorm, RxTerms, and NDF-RT were found to be pre-coordinated representations of dose form, administration, and/or delivery device. Semantic decomposition and concept grouping identified 9 distinct properties that were included in the original representations of dose form (Table 1). Three of the properties were quantities that could be represented using the Unified Code for Units of Measure (UCUM). The other 6 properties represented coded data (CD); value sets for these properties were constructed using concepts derived from 5 sources (RxNorm, RxTerms, NDF-RT, USP Forum, and the FDA standard). With few exceptions, each of the original terms could be represented using the extended properties and associated value sets.

Discussion and Conclusion

As previously observed[1], the existing terms for dose form were highly pre-coordinated, which complicated the use of those terms for medication CDS. We decomposed those terms into semantically "pure" concepts and grouped them into 9 distinct properties. Value sets were derived from existing authoritative and standards-based resources.

The proposed properties and their corresponding value sets were able to represent nearly all of the concepts in the original pre-coordinated terms. The proposed terms and mappings are currently under review by pharmacists to validate their semantic accuracy and completeness. These proposed extensions would significantly improve the ability to create and maintain drug lists for medication CDS.

Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomics Research Network) and the Mayo Clinic Office of Information and Knowledge Management.

Property	Datatype	Primary Source of Terms	Examples
Dose Form			
Drug Form	CD	NDF-RT Orderable Drug Form (modified)	Aerosol, Lotion, Capsule
Delivery Form	CD	NDF-RT Orderable Drug Form (modified)	Aerosol, Lotion, Capsule
Release Pattern	CD	U.S. Pharmacopeial Forum	Immediate, Extended
Extended Release Time	QTY	UCUM	24 HR (hour)
Administration			
Route	CD	SNOMED-CT Route Subset	Buccal, Intramuscular
Method	CD	NDF-RT, RxTerms (derived)	Drops, Inhalation, Injection
Delivery Device			
Type	CD	NDF-RT Drug Delivery Device (modified)	Drug Eluting Implant (Rod), Metered Dose Inhaler, Patch
Capacity	QTY	UCUM	1 mL
Extended Release Time	QTY	UCUM	21 D (day)

Table 1: Proposed extensions for RxNorm

References

1. Freimuth RR , Wix K, Zhu Q, Siska M, Chute CG. Evaluation of RxNorm for Medication Clinical Decision Support. AMIA 2014 Annual Symposium. Nov 2014.

Evaluating the Use of Star Allele Nomenclature with High-Throughput Sequence Data: Implications for Research and Clinical Practice

Adam S. Gordon[1]; Deborah A. Nickerson, PhD[1]; Christopher G. Chute, MD, DrPH[2]; Robert R. Freimuth, PhD[3,4]

[1]Department of Genome Sciences, University of Washington, Seattle, WA

[2]Division of General Internal Medicine, Johns Hopkins University, Baltimore, MD

[3]Department of Health Sciences Research and [4]Office of Information and Knowledge Management, Mayo Clinic, Rochester, MN

Summary

We developed a novel algorithm to evaluate the impact of the classical "star allele" nomenclature system when it is used with exome sequencing data. We catalogued and quantified the errors identified in 5 genes from 6503 individuals, many of which could negatively impact research studies and clinical practice.

Introduction and Background

The number of clinically-implemented pharmacogenomics (PGx) tests is growing rapidly. Most PGx genes use a "star allele" nomenclature system to describe known genetic variations, and this system is widely used to represent genotype. In particular, much of the clinical decision support (CDS) that has been implemented for PGx genes uses star alleles as the basis for recommending a course of action to the provider. While the robust representation of test results is essential for deriving accurate clinical interpretations and recommendations, the ability of the star allele nomenclature system to represent next-generation sequencing (NGS) results has not been systematically evaluated. Limitations in nomenclature will negatively impact data quality and patient care.

Materials and Methods

To quantitatively analyze the ability of star nomenclature systems to accurately represent NGS results, we developed a novel algorithm that uses curated versions of public allele definition tables to translate phased, NGS-derived genotypes (in VCF format) into star allele-based diplotypes. Phased exome sequence data from 6503 individuals of European-American or African-American ancestry, drawn from the NHLBI Exome Sequencing Project, was used to simulate a large cohort of clinical exome data. We cataloged the type and frequency of errors encountered in 5 genes with clinically actionable alleles (*CYP2C9*, *CYP2C19, CYP3A5, SLCO1B1,* and *TPMT*).

Results

The accuracy of the star nomenclature system varied widely across the 5 genes. Errors in allele assignment were due to rare variants, potential phasing errors, and partial or mixed haplotypes. Some errors had clear clinical impact. For example, in *CYP2C9*, which is used to inform dosing of warfarin, 25.3% of African-Americans and 14.6% of European-Americans in this dataset carry at least one allele that cannot be named using the canonical star allele definitions. As a result, if only those alleles regularly interrogated by clinical labs were considered, nearly 48% of African-Americans and 16% of European-Americans would be assigned an incorrect diplotype. These results demonstrate the need for more accurate systems for reporting clinical NGS results.

Discussion

As clinical genetic testing moves from genotyping to NGS-based platforms, there is a growing need to quantitatively assess the utility of classical data representation schemes, including star allele nomenclature systems. The results of this study reveal that careful consideration must be taken when using legacy star nomenclature systems to report NGS data, especially for under-sequenced populations. This innovative analysis is, to our knowledge, the first effort to quantify naming errors using large-scale, individual level data and to estimate their impact on clinical practice.

Acknowledgment

This work was supported by the NIH/NIGMS (U19 GM61388; the Pharmacogenomic Research Network).

Search Tag Analyze Resource (STAR): An online platform to crowd-source genomic disease signatures from open digital samples.

Dexter Hadley[1], MD/PhD; James Pan[1]; Marina Sirota[1], PhD, Bin Chen[1], PhD; Boris Oskotsky[1], PhD; Raymond K. Auerbach, PhD; Alexander Morgan, PhD; Benjamin Pinsky[3], MD/PhD; Atul J. Butte[1, 2], MD/PhD

[1]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305 USA
[2]Lucile Packard Children's Hospital, Palo Alto, CA 94304 USA
[3]Stanford University Medical Center, Department of Pathology and Laboratory Medicine, Stanford, CA 94305 USA

Abstract

Despite the continued exponential growth of genomic data in the public domain, few resources facilitate shared annotation and analysis of open digital samples. The sheer volume of public data invites large-scale collaborative interpretation by experts from the scientific community; however, a significant bioinformatics burden exists for most physicians and scientists to download, curate, and analyze digital samples. To this end, we designed STAR, the Search Tag Analyze Resource (http://stargeo.org), to crowd-source curation and analysis of open digital samples from NCBI's Gene Expression Omnibus (GEO). The online platform provides search, annotation, and analytical engines that allow users to design and execute powerful meta-analytics across digital samples in the cloud. As a proof of concept, we used STAR to design and conduct a global experiment to investigate severe dengue susceptibility by identifying significant robust gene signature meta-effects in 197 open digital samples with severe dengue relative to 353 digital samples with uncomplicated dengue. Samples were drawn across eight experiments from populations in Brazil, Cambodia, Malaysia, Nicaragua, Singapore, Thailand, Venezuela, and Vietnam. Among our leading 10 targets from STAR are nine genes encoding proteins that are upregulated and enriched in neutrophils such as *LTF*, *CEACAM8*, *PGLYRP1*, and *ELANE* among others. Moreover, these nine genes highlight a common pro-inflammatory cytokine pathway mediated by TNFα, a central pathway to the pathophysiology of severe dengue and a usual target of many drugs. Therefore, our leading candidates are ideal targets to develop as prognostic biomarkers to predict the clinical course of dengue infection, and to prioritize treatment for those most at risk of severe complications. Our results prove the translational utility of STAR as an online platform to identify and annotate open digital samples to meta analyze for genomic disease signatures. Furthermore, we posit that collaboratively mining public genomics data repositories may fuel novel biomarker and drug discoveries of significant translational impact and empower physicians and scientists to molecularly redefine the meaning of disease.

Introduction and Background

The human genome comprises tens of thousands of genes that are functionally disrupted in disease. Since their development over two decades ago, gene expression microarrays have facilitated the identification of gene signatures that molecularly characterize disease at the functional level. Moreover, many "digital samples" assayed by microarray continue to be deposited into public data repositories at an unprecedented scale. To encourage data sharing and collaboration among the scientific community, the NCBI's Gene Expression Omnibus (GEO) openly shares over 1.2 million digital samples across 50,000 functional genomics experiments that have been deposited to-date. The sheer volume of functional data currently in the public domain invites collaboration on both the analysis and interpretation of digital samples among many specialized expert physicians and scientists that are familiar with the disease.

Meta-analysis refers to powerful statistical methods for combining results of individual studies to find robust effects with support across studies. This method has long been used on public data to inform clinical guidelines and decisions. We and others have been meta-analyzing genome-wide microarray experiments from GEO and other public data repositories to define robust gene signatures that we

translate into novel biomarkers and therapeutics for disease. With this approach, we have successfully developed novel biomarkers and / or therapeutic strategies for organ transplant rejection [1], lung cancer [2], pancreatic cancer [3], chronic renal disease [4], and preeclampsia [5,6] among others. In our experience, defining robust gene signatures from public data involves a computationally laborious process requiring substantial technical expertise to download, curate, and meta-analyze digital samples. This significant bioinformatics burden represents a major bottleneck to collaborative analysis and interpretation of open functional genomic data because it disconnects a majority of researchers that may have invaluable disease expertise to contribute, but who lack a computational background to properly mine the public data.

Furthermore, while as a community we continue to share digital samples through GEO and other public data repositories[7–9], in general, the data remains poorly annotated for secondary in depth biomedical research. In order to maximize data submission, public data repositories generally do not impose any biologically interpretable schema/ontology on the deposited digital samples. Thus, few resources facilitate the shared analysis and biological interpretation of "raw" public functional genomic data. For instance, of the over 1.2 million samples deposited in GEO it is unclear how many of them have a specified disease status on a per sample level. Therefore, as public data repositories continue to grow exponentially, the lack of interpretable biological annotations of digital samples represents a worsening problem for the functional genomics community.

In this work we describe the Search Tag Analyze Resource (STAR), an online platform to enable collaborative analysis and interpretation of shared digital samples. STAR represents an annotation layer built on open public data, and it provides convenient tools that remove the substantial bioinformatics burden currently required to derive robust gene signatures from public functional genomics data. In essence, STAR is a web application to enable users to:

1) Search full text sample level attributes within public data repositories
2) Tag digital samples with biologically relevant annotations (for example age, gender, ethnicity, or disease), and
3) Analyze tagged and annotated samples across different studies with meta-statistics to define robust gene signatures effects for disease

We hypothesize that by providing STAR to the global functional genomics community as an open online platform, we can comprehensively crowd-source and scale the annotation and analysis of digital samples in the public domain. Moreover, we posit that collaboratively mined robust disease gene signatures will fuel novel biomarker and drug discoveries of significant translational impact. Furthermore, accurately representing different diseases in terms of their functional genomic signatures enables physicians and scientists to start redefining the very nature of disease itself.

As a proof of concept, we used STAR to search GEO for gene signatures that predict severe dengue. Dengue is a mosquito-borne viral infection causing a flu-like, febrile illness that can progress to severe dengue. Severe dengue is characterized by a constellation of symptoms that include lethal complications of dengue hemorrhagic fever and dengue shock. Dengue is the most important arthropod-borne viral infection of humans in the world with over 3.6 billion people at risk of contracting 2 million cases of severe dengue causing an estimated 21,000 fatalities yearly. Moreover, severe dengue is a leading cause of serious illness and death among children in Asian and Latin American developing countries. Although there are no specific drugs for severe dengue, administering fluids and electrolytes early in the clinical course lowers fatality rates from 20% to below 1%.

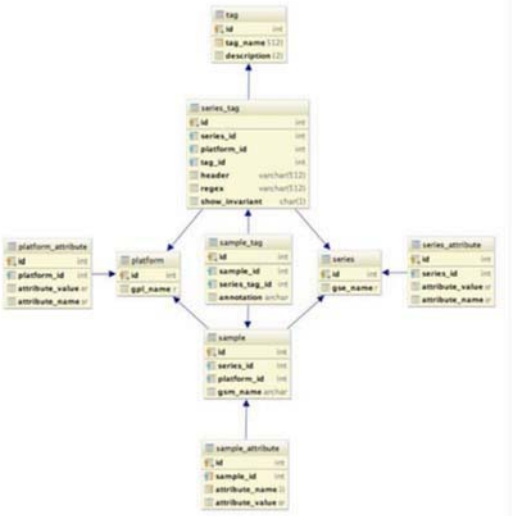


Figure 1: The STAR schema. The figure shows the database entity relationships centered on Sample Tags, which hold biological annotations of digital samples.

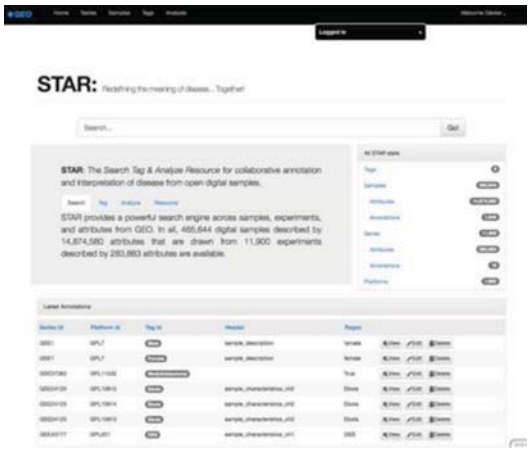


Figure 2: The STAR online portal. The figure shows a screen capture of http://stargeo.org.

designed a tagging and annotation interface for users to use regular expressions to quickly annotate user defined sample tags with biological interpretation. Additionally, we provide a simple analytical interface for users to design, compute and visualize standard genetic meta-analysis of random and fixed effects across tagged and annotated digital samples. Finally, we integrate proven social online technologies such as wikis and ratings to foster collaborative interpretation.

Despite being highly treatable, epidemic outbreaks of dengue can quickly overwhelm the resource-constrained health systems of developing countries where the disease is endemic. However, fatalities can be minimized if life-saving supportive care is prioritized early for those patients most at risk of developing severe complications. Since complications typically develop two weeks after the onset of fever, we used STAR to design and execute a meta-analysis to define a robust gene signature for severe dengue that could predict the onset of hemorrhagic fever and shock syndrome in patients within the first week of fever onset. Such gene signatures predictive of severe dengue can be translated into prognostic biomarkers to be used by clinicians to significantly improve disease morbidity and mortality.

Methods

Using the Amazon Web Services cloud infrastructure, we downloaded over 1.7 TB of public data for all processed expression data and associated attributes for series, samples, and platforms catalogued in GEO (ftp://ftp.ncbi.nih.gov/pub/geo/DATA/), and we developed a scalable database schema to represent their attributes. We designed the schema around sample tags (Figure 1), which are user-assigned key:value bindings for digital samples where key sample tags are bound to sample annotation values (for example Age:50, Gender:Female, Cancer:True, etc.) on an open-source PostgreSQL relational database management system backend. With this schema, we implement a semantic network to crowd-source tags that represent biological annotations.

For the prototype web application described here (Figure 2), we filtered for “expression profiling by microarray” in humans to find 465,644 digital samples from 11,900 series (experiments) across 1,682 different platforms. We full text indexed all 14,874,580 sample and 283,883 series attributes to facilitate rapid searches at the sample attribute level, a task currently impossible on GEO. Moreover, we

Table 1: Digital sample inventory. The table lists the eight studies from GEO and counts of samples annotated for normal uncomplicated dengue (DF=dengue fever) and severe dengue complications (DHF = dengue hemorrhagic fever, DSS = dengue shock syndrome).

Study	Country	Normal	Severe			Grand
		DF	DHF	DSS	Total	Total
GSE13052	Vietnam	9		9	9	18
GSE17924	Cambodia	16	13	19	32	48
GSE18090	Brazil	8	10		10	18
GSE25001	Vietnam	89		37	37	126
GSE25226	Nicaragua	20	6	8	14	34
GSE38246	Nicaragua	50	26	19	45	95
GSE40628	Vietnam	6	6	1	7	13
GSE43777	Venezuela	154	43		43	197
Grand Total		352	197	104	93	549

Results

We used STAR to search series attributes from GEO for the "dengue" keyword to find 20 series containing 1540 digital samples that we tagged as "DF" (Dengue Fever), "DHF" (Dengue Hemorrhagic Fever), "DSS" (Dengue Shock Syndrome), and "Acute" (acute infection ≤ 7 days of illness). Furthermore, we used STAR to meta-analyze differential gene expression in "Acute" samples that go on to develop severe dengue ("DHF"|"DSS") relative to those with only uncomplicated dengue ("DF"). In all, we defined a robust gene signature across 8 experiments using 197 digital samples with severe dengue in comparison to 353 digital samples with uncomplicated dengue (Table 1). These well-curated samples were drawn from eight studies performed in Brazil, Cambodia, Malaysia, Nicaragua, Singapore, Thailand, Venezuela, and Vietnam, and they were all collected from dengue patients that presented in the acute phase of infection (i.e. within one week of fever onset).

Our leading 10 gene targets by effect size (Table 2) are mainly upregulated and enriched in neutrophils encoding proteins such as Lactotransferrin (*LTF)*, Carcinoembryonic Antigen-Related Cell Adhesion Molecule 8 (*CEACAM8*), Peptidoglycan Recognition Protein 1 (*PGLYRP1*), and Neutrophil Elastase (*ELANE*) among others. We found nine out of 10 genes of most significant effect highlight a common pro-inflammatory cytokine pathway mediated by TNFα (Figure 3), a central pathway in the pathophysiology of severe dengue[10], and a usual target of many drugs. Therefore, many of these candidates represent ideal targets to developed as prognostic biomarkers to predict the clinical course of dengue infection, and to prioritize treatment for those most at risk of severe complications. Indeed, a literature search revealed that neutrophil elastase, *ELANE*, has already been validated as a successful biomarker for severe dengue in children[11] in at least one study, and our results for *ELANE* (Figure 4) confirm this finding with replication on an international scale.

Discussion

In this work, we describe STAR, the Search Tag Analyze Resource, to crowd-source annotation and analysis of open digital samples. The online platform provides search, annotation, and analytical engines that allow users to design and

Table 2: Top 10 genes ranked by significant effect. The table lists the gene, effect size (log2 fold change), uncorrected p value, the number of cases and controls, and the number of different studies.

Gene	Effect	p	#case	#ctrl	k
LTF	0.94	4.00×10^{-3}	138	282	7
CEACAM8	0.93	4.72×10^{-4}	131	276	6
PGLYRP1	0.84	1.87×10^{-3}	131	276	6
DEFA4	0.83	4.38×10^{-3}	138	282	7
ELANE	0.81	5.69×10^{-4}	131	276	6
MMP8	0.74	4.11×10^{-5}	131	276	6
CEACAM6	0.59	1.38×10^{-3}	138	282	7
CTSG	0.55	2.15×10^{-3}	138	282	7
HLA-DQB1	-0.55	7.15×10^{-3}	150	296	8
ARG1	0.53	1.59×10^{-4}	138	282	7

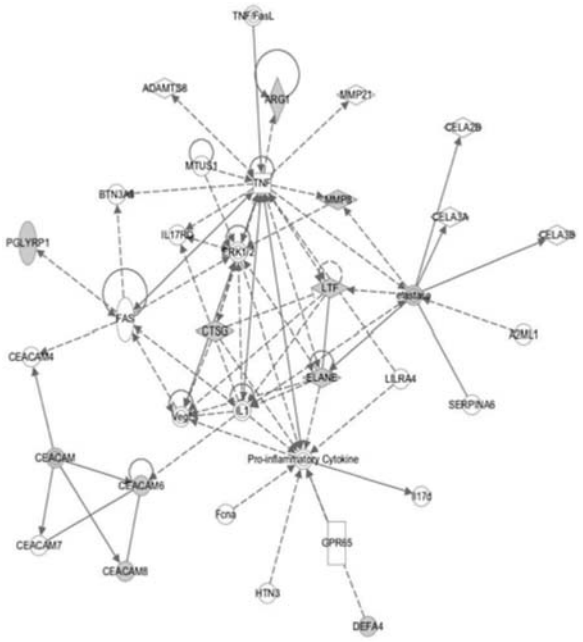


Figure 3: Common pro-inflammatory cytokine pathway mediated by TNFα is dysregulated in severe dengue. The figure shows results from Ingenuity Pathway Analysis where nine out of 10 genes of most significant effect highlight this shared pathway.

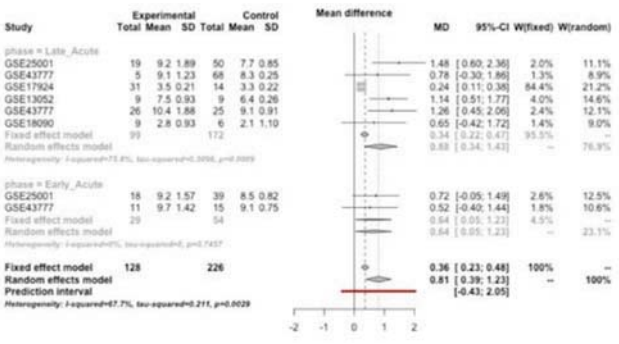


Figure 4: Meta-analysis of *ELANE*. The figure shows a Forrest plot of six experiments that assayed *ELANE*. Early and Late Acute phases refer to 0-3 days and 4-7 days respectively since fever onset.

execute powerful meta-analysis across digital samples in the cloud. We demonstrate its translational utility in a case study where we identify a robust gene signature using eight different microarray studies from around the world that implicate both validated biomarkers like *ELANE* for severe dengue as well as defective common pro-inflammatory cytokine pathway mediated by TNFα which is central to the pathophysiology severe dengue susceptibility.

We provide STAR to the larger scientific community to democratize and scale the interpretation of public functional genomics data. Indeed, an international study of this scale on a neglected tropical disease such as dengue is impossible without open public data. Also, STAR's cloud infrastructure improves speed, convenience, and accuracy in mining open digital samples. But most importantly, STAR allows for a boundless amount of global talent to collectively interpret an endless supply of open data. In our experience mining the public data, it is physicians and scientists most familiar with disease that propose the most impactful and translational hypotheses to test. We make STAR available so that such specialists can test their hypotheses by designing powerful meta-analytical studies across hundreds of thousands of digital samples without any bioinformatics training or experience.

Using STAR we were able to find validated biomarkers for severe dengue and a dysfunctional TNFα-regulated pro-inflammatory cytokine pathway in less than one day using only open data. We envision that the platform will gain utility as it grows and we crowd-source additional biological tags. We posit that collaboratively mining public genomics data repositories may fuel novel biomarker and drug discoveries of significant translational impact, and we present STAR as an open tool to ultimately empower physicians and scientists to molecularly redefine the meaning of disease.

Conclusion

Our results prove the translational utility of STAR as an online platform to define genomic disease signatures from open digital samples.

Acknowledgements

We thank Boris Oskotsky from Stanford University for computer support. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01 GM079719, and the National Institute of Allergy and Infectious Diseases (Bioinformatics Support Contract HHSN272201200028C). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Chen, R. *et al.* Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions. *PLoS Comput. Biol.* **6,** (2010).
2. Vicent, S. *et al.* Cross-species functional analysis of cancer-associated fibroblasts identifies a critical role for CLCF1 and IL-6 in non-small cell lung cancer in vivo. *Cancer Res.* **72,** 5744–56 (2012).
3. Sharaf, R. N. *et al.* Computational prediction and experimental validation associating FABP-1 and pancreatic adenocarcinoma with diabetes. *BMC Gastroenterol.* **11,** 5 (2011).
4. Butte, A., Sigdel, T. & Wadia, P. Protein microarrays discover angiotensinogen and PRKRIP1 as novel targets for autoantibodies in chronic renal disease. *Mol. Cell. …* (2011). at <http://www.mcponline.org/content/10/3/M110.000497.short>
5. Wen, Q. *et al.* Peptidomic Identification of Serum Peptides Diagnosing Preeclampsia. *PLoS One* **8,** e65571 (2013).
6. Liu, L. Y. *et al.* Integrating multiple "omics" analyses identifies serological protein biomarkers for preeclampsia. *BMC Med.* **11,** 236 (2013).
7. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39,** 1181–6 (2007).
8. Bhattacharya, S. *et al.* ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* **58,** 234–9 (2014).
9. Rustici, G. *et al.* ArrayExpress update--trends in database growth and links to data analysis tools. *Nucleic Acids Res.* **41,** D987–90 (2013).
10. Clyde, K., Kyle, J. L. & Harris, E. Recent advances in deciphering viral and host determinants of dengue virus replication and pathogenesis. *J. Virol.* **80,** 11418–31 (2006).
11. Juffrie, M. *et al.* Inflammatory mediators in dengue virus infection in children: interleukin-8 and its relationship to neutropzhil degranulation. *Infect. Immun.* **68,** 702–7 (2000).

A systems biology approach to genome-environment interactions using zebrafish (*Danio rerio*)

Gary Hardiman, Department of Medicine, Medical University of South Carolina

Endocrine disrupting chemicals (EDCs) including plasticizers, pesticides, detergents and pharmaceuticals, affect a variety of hormone-regulated physiological pathways in humans and wildlife. Many EDCs are lipophilic molecules and bind to hydrophobic pockets in steroid receptors, such as the estrogen receptor and androgen receptor, which are important in vertebrate reproduction and development. Indeed, health effects attributed to EDCs include reproductive dysfunction (e.g., reduced fertility, reproductive tract abnormalities and skewed male/female sex ratios in fish), early puberty, various cancers and obesity. A major concern is the effects of exposure to environmentally relevant concentrations of endocrine disruptors in utero and post partum, which may increase the incidence of cancer and diabetes in adults. EDCs affect transcription of hundreds and even thousands of genes, which has created the need for new tools to monitor the global effects of EDCs.

The zebrafish [*Danio rerio*] has emerged as an important tool for studying the biological effects of hormones and EDCs. Zebrafish is a small tropical fresh-water fish indigenous to Asia. Compared with other model organisms (i.e. mice and rats) zebrafish are relatively inexpensive to maintain. They are oviparous and are easily bred in large numbers. Zebrafish have short life spans [60 days to maturity]. Thus, three generations of zebrafish can be easily generated in a year, making it possible to carry out immediate studies in response to exposure to endocrine disruptors, in addition to longitudinal transgenerational studies analysis with progeny derived from the exposed fish. Presently there is no marine vertebrate model that is as well characterized as the zebrafish. Their highly inbred, lab-dependent nature and well defined genome make them an ideal model for toxicology studies and rapid phenotypic assessment.
Zebrafish conserve many developmental pathways found in humans, which makes zebrafish a valuable model system for studying EDCs especially on early organ development because their embryos are translucent.

A major technological shift in the biomedical research community over the past decade has been the adoption of high throughput or massive parallel sequencing (HTS) technologies to facilitate whole genome and transcriptome sequencing. Application of HTS for investigating gene transcription provides a sensitive tool for monitoring the effects of EDCs on humans and other vertebrates as well as elucidating the mechanism of action of EDCs.

In this presentation recent advances in massive parallel sequencing approaches will be reviewed with a focus on zebrafish in the context of genome-environment interactions. Zebrafish exposed to EDCs at different stages of development, can provide important insights on EDC effects on human health. I will describe an unbiased, unsupervised discovery approach to elucidate the intermediate, elementary interactions that occur when key transcriptional processes are disrupted during early development. Xenoestrogen exposures that link to human disease phenotypes can be determined in this manner and the key signaling pathways identified. The zebrafish developmental model can thus serve as a proxy for human health assessment to study the effect of xenoestrogens that link to human disease phenotypes.

Resources from the Clinical Pharmacogenetic Implementation Consortium (CPIC) to Enable Pharmacogenomic Clinical Decision Support

James M. Hoffman, PharmD;[1] Michelle Whirl-Carrillo, PhD;[2] Robert R. Freimuth, PhD;[3] Cyrine E. Haidar, PharmD;[1] Kelly E. Caudle, PharmD;[1] Teri E. Klein, PhD;[2] Mary V. Relling, PharmD[1]
[1]St. Jude Children's Research Hospital, Memphis, TN, USA; [2]Stanford University, Palo Alto, CA, USA; [3]Mayo Clinic, Rochester, MN, USA

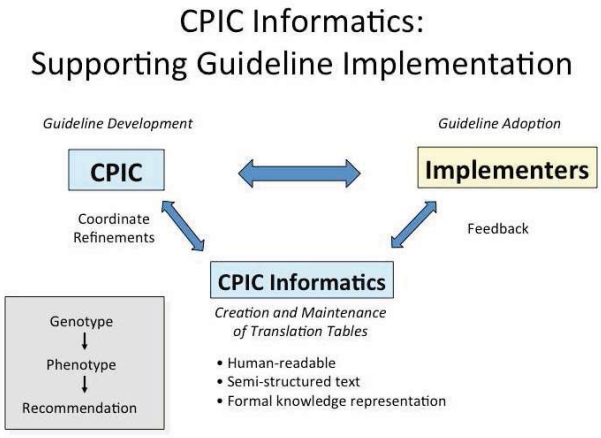
Summary: Clinical decision support (CDS) is essential to the implementation of pharmacogenomics. The Clinical Pharmacogenetic Implementation Consortium (CPIC) creates comprehensive tables and other guidance in each new and updated guideline to translate genotype information to phenotype to clinical recommendation, using human readable and structured text with formal knowledge representation.

Background: Clinical decision support (CDS) is essential to the implementation of pharmacogenomics. Unlike other types of medication CDS where vendors provide a standard set of CDS functions that are often refined for site specific requirements, pharmacogenomic CDS requires each organization to expend substantial effort to design all aspects of the CDS. A primary reason for a site specific approach is the lack of a curated and machine-readable database of pharmacogenomic knowledge suitable for use in an electronic health record (EHR) with CDS.

Methods: CPIC develops detailed gene/drug clinical practice guidelines, which enable the translation of genetic laboratory test results into actionable prescribing decisions for specific drugs. CPIC guidelines are standardized, including a uniform system for grading levels of evidence linking genotypes to phenotypes, assigning phenotypes to clinical genotypes, prescribing recommendations based on genotype/phenotype, and assigning a strength to each prescribing recommendation. Each guideline is peer-reviewed and freely available on PharmGKB (https://www.pharmgkb.org/page/cpic). Recognizing the potential to facilitate the adoption of CPIC guidelines in EHRs with CDS, a formal working group was established within CPIC in 2013 to facilitate the adoption of the CPIC guidelines by identifying, and resolving where possible, potential technical barriers to the implementation of the guidelines within a clinical electronic environment. CPIC creates comprehensive tables and other guidance to translate genotype information to phenotype to clinical recommendation. CPIC resources are currently available using human readable and structured text.

Results: Each new and updated CPIC guideline includes vendor agnostic resources to enable pharmacogenomic CDS. Figures outline workflow processes for EHR implementation point of care CDS. Tables are provided to translate genotype results into interpreted phenotypes, diplotype/phenotype interpretative summaries for each possible diplotype combination for variants with known functional significance, and example CDS text. A more comprehensive table is also posted to PharmGKB, which summarizes each possible diplotype combination for all the variants listed in the guideline supplement, including "possible" phenotypes to account for ambiguities in diplotype interpretation (e.g., possible poor metabolizer). These resources are standardized in each guideline, but flexibility remains to accommodate more complex situations, such as when a drug (e.g. phenytoin) requires 2 genes (e.g, *CYP2C9* and/or *HLA-B*) for interpretation.

Discussion: CPIC provides unique resources to enable pharmacogenomic CDS in any EHR, which will be refined based on feedback from implementers (Figure). The knowledge framework developed by the Clinical Decision Support Consortium (CDSC) serves as a reference point for standardizing informatics content in CPIC guidelines. While current content is focused on Level 1 (unstructured) and Level 2 (semi-structured) artifacts, opportunities for higher level knowledge representations are being evaluated.



Meta-analysis of Gene Expression Using the Elastic Net

Jacob J. Hughey[1], Atul J. Butte[1]

[1] Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA

Summary

Meta-analysis of gene expression has led to numerous insights, but current methods have several limitations. To address these limitations, we developed a method to perform meta-analysis of gene expression using the elastic net. Our method will be a valuable tool for extracting knowledge from publicly available gene expression data.

Introduction

Meta-analysis of gene expression has led to numerous insights into biological systems, but current methods have several limitations. To address these limitations, we developed a method to perform meta-analysis of gene expression using the elastic net. The elastic net, a generalization of ridge and lasso regression, is a powerful and versatile method for classification and regression. Using our method, we performed a meta-analysis of lung cancer.

Methods

Each dataset is normalized using standard techniques. ComBat is used to correct for batch effects. Glmnet (R implementation of the elastic net) is used to perform cross-validation and to train and test a predictive model.

For our meta-analysis of lung cancer, we used five datasets for discovery and three for validation. All are publicly available. In each dataset, we used only the samples that were histologically defined as adenocarcinoma (AD), squamous cell carcinoma (SQ), small cell lung carcinoma (SCLC), or carcinoid (CAR).

Results

Based on leave-one-study-out cross-validation (Figure 1), the optimal classifier for distinguishing the four types of lung cancer contained 59 genes. Almost every gene in the classifier was informative (i.e., had a non-zero coefficient) for only one type of lung cancer. On three independent datasets, the classifier showed an overall accuracy of 91%.

Discussion

Our approach expands the types of questions that can be addressed using meta-analysis of gene expression. We believe our method will be a valuable tool for extracting knowledge from the steadily increasing amount of publicly available gene expression data.

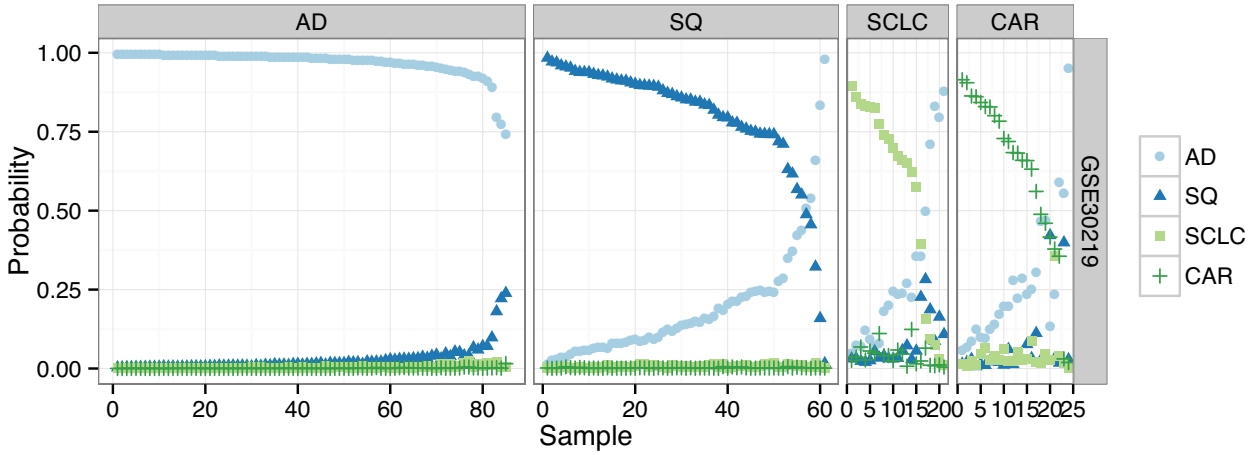


Figure 1. Estimated class probabilities for each sample in one study, based on cross-validation. The title of each panel denotes the "true" class of the samples in that panel. For each sample, there are four points, corresponding to the estimated probability that the sample belongs to each class. Within each panel, samples are sorted by the probability of the true class. For example, all samples in the left-most panel have a high probability of AD, which means they were classified correctly.

Partial integration strategy of heterogeneous datasets to prognosticate survival in glioblastoma

Haruka Itakura, MD, MS, Olivier Gevaert, PhD
Division of Biomedical Informatics, Dept of Medicine, Stanford University, Stanford, CA

Abstract

We developed a data integration strategy that integrates heterogeneous datasets to prognosticate survival in glioblastoma. In contrast to simple concatenation of feature vectors, our integration strategy was able to identify a profile set of features from both datasets that determined differential survival.

Introduction

Although the most commonly used technique to integrate data is simple concatenation of feature vectors, it is not the most accurate and often yields results that are no better than results from single datasets alone [1-3]. We sought to develop a data integration strategy that integrates heterogeneous datasets to prognosticate survival in glioblastoma (GBM), a highly lethal and the most common primary malignant brain tumor.

Methods

We obtained survival, gene expression, and methylation status data from TCGA on subjects with GBM. Gene expression data underwent processing by a previously described method[4] and was divided into a group of 100 modules that clustered around central driver genes. Taking the subset with the intersection of these datasets we fit a Cox regression model using penalized maximum likelihood. We trained and tested our Cox proportional hazards prediction model using 10-fold cross-validation and also used internal cross-validation to minimize the regularization parameter lambda. We performed the regression analyses on individual datasets, a concatenated dataset, and a partially integrated dataset. In the partially integrated dataset, we first identified features that had statistically significant beta coefficients in individual models, then included them in subsequent models when predicting hazards ratios in Cox regression. All data were normalized using z-score standardization prior to analysis.

Results

Survival data were available on 575 subjects, gene expression module data of 100 features on 426, and methylation status of 348 genetic markers on 402. 255 subjects, who possessed all datasets, were included for analysis. Running the gene expression module dataset alone yielded a multivariate combination of three modules ($p=0.079$). Of these three modules one was favorable and two unfavorable for survival. When running the methylation status dataset alone, a multivariate combination of 12 features emerged as statistically significantly contributing to survival ($p=0.011$); six of these features were favorable for survival, five were not. Concatenation of gene expression module and methylation data revealed a combination of three features favorably affecting survival and two features negatively affecting survival ($p=0.063$). Four features were significant in the individual methylation and module status models. Performing partial integration we identified a multivariate combination of five features that contributed to survival ($p=0.0053$), where two related to methylation status and were favorable for survival and three were gene expression modules, of which two were unfavorable for survival.

Conclusion

Whereas individual datasets produced features that prognosticated survival in GBM, integrating these datasets using simple concatenation eliminated any predictive signals. Our partial integration strategy, however, generated a multivariate combination of five features from two heterogeneous datasets with statistical significance ($p=0.0053$). These findings suggest the effectiveness of our integration strategy, which aimed to overcome the limitations of the most commonly-used concatenation method.

References

1. Gevaert O, De Smet F, Timmerman D, Moreau Y, De Moor B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 2006; 22(14): E184-190.
2. Daeman A, Gevaert O, De Moor B. Integration of clinical and microarray data with kernel methods. Engineering in Medicine and Biology Society, 2007 EMBS 2007 29th Annual International Conference of the IEEE; 2007 22-26 August. 2007. P.5411-5.
3. Boulesteix AL, Porzelius C, Daumer M. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics 2008; 24: 1698-1706.
4. Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. Interface focus 2013; 3:20130013.

Network-Based Models of Context-Specific Synthetic Lethality

Alexandra Jacunski[1], MA; Nicholas Tatonetti, PhD[2-4]
[1]Integated Program in Cellular, Molecular, and Biomedical Studies; [2-4]Departments of Biomedical Informatics, Systems Biology, and Medicine, Columbia University, New York, NY

Summary: Synthetic lethality (SL) is a genetic interaction with many applications to human health, including identifying cancer combination therapies. Research suggests that SL interactions may differ between environments.(1) Here, we predict context-specific SL by augmenting our SINaTRA algorithm (Species-Idependent Network TRAnslation; in submission), which models SL between species, with DaVIS (Distinguishing enVironment-specific InteractionS).

Introduction: Network medicine contextualizes systems using mathematical graphs and computational methods; for example, it depicts the protein-protein interaction (PPI) network by using genes as nodes, and joining two nodes if the gene products physically interact. The PPI network has previously been used to create a computational model of SL, a genetic interaction where two nonessential genes cause cellular or organism inviability if knocked out simultaneously. We previously used SINaTRA to create translational models of SL by normalizing the network of a source species to predict SL in a target species. For example, a model trained on the *S. cerevisiae* PPI network was able to predict SL using the *S. pombe* network (Figure 1A).

Recent research indicates that PPIs rewire under stress,(2) and SL changes in different environments.(1) We hypothesized that we could use weighted *S.cerevisiae* PPIs in control conditions as our source to predict SL in *S. cerevisiae* under stress as the target. We created context-specific PPI networks using environment-specific gene coexpression data. This method can identify tumour-, disease-, or tissue-specific SL pairs.

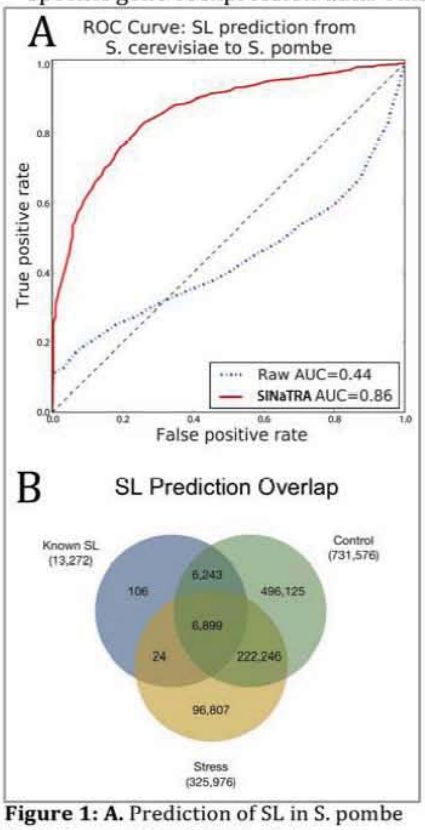


Figure 1: A. Prediction of SL in S. pombe from S. cerevisiae; blue depicts untranslated network parameters, while red indicates performance after translation. **B.** Identification of true SL pairs (blue) in S. cerevisiae using the expression-weighted control (green) and stress (yellow) PPI networks.

Methods: We constructed single-component protein-protein interaction (PPI) networks for *S. cerevisiae* using experimental data from BioGrid. We weighted *S. cerevisiae* network edges using context-specific gene coexpression values to create two context-specific PPI networks: control and peroxide-stressed (microarray data from GEO DataSets). We then calculated the 14 network parameters used in our original SINaTRA algorithm, which uses normalized parameter values to train the model. Parameter algorithms accounted for weighted edges where applicable. We trained our model using SL data downloaded from BioGrid on the control network, and applied this model to the peroxide-stressed network. Using this method, we were able to detect stress-specific SL pairs.

Results: In previous work (in submission), we used the SINaTRA algorithm to create a species-agnostic model in order to predict SL in species where the interaction is understudied. We expanded on these findings by identifying stress-specific SL pairs. We found that, while the control network correctly predicted 13,142 out of the 13,272 experimentally identified SL pairs, only 6,923 of the original pairs are expected to remain SL in the peroxide-stressed network (Figure 1B). We also identified context-specific SL in heat- and starvation-stressed networks.

Discussion: Our results indicate that expression-weighted networks are able to predict SL in *S. cerevisiae* under control conditions. The control network can be used to predict SL in other environments, such as peroxide stress, in turn. Since experimental elucidation of human SL must occur in cell lines, this methodology may be applied to confirming whether pairs that are SL in one cell line, with its specific expression patterns, will hold in others. Future work will focus on experimental validation and extrapolation to human cancers.

References

1. Chan N, Pires IM, Bencokova Z, Coackley C, Luoto KR, Bhogal N, et al. Contextual Synthetic Lethality of Cancer Cell Kill Based on the Tumor Microenvironment. Cancer Research. 2010 Oct 13;70(20):8045–54.

2. Lehtinen S, Marsellach FX, Codlin S, Schmidt A, Clément-Ziza M, Beyer A, et al. Stress induces remodelling of yeast interaction and co-expression networks. Mol BioSyst. 2013;9(7):1697.

medTurk: an In-house Crowdsourcing Approach to Extracting Information from Clinical Notes

Robert M. Johnson, *MS*[1]; Shruti Rao, MS[1]; Priya Kasturirangan, MBBS[2]; Aziza T. Shad, MD[2]; Kenneth P. Tercyak, PhD[3]; Subha Madhavan, PhD[1]
[1]Innovation Center for Biomedical Informatics; [2]Division of Pediatric Hematology Oncology, Blood and Marrow Transplantation; [3]Division of Population Sciences Lombardi Comprehensive Cancer Center

Summary

We developed medTurk, a software prototype modeled after Amazon's Mechanical Turk that coordinates in-house crowdsourcing to reliably and confidentially extract information from clinical notes. We demonstrate an application of this prototype to extract late effects among childhood cancer survivors from clinical notes.

Introduction and Background

Most of the clinical information in EHRs is present in the form of narrative reports. Although there exist several natural language processing and machine learning algorithms that perform and support specific information extraction tasks, such as cTAKES (clinical Text Analysis and Knowledge Extraction System), these automated techniques cannot guarantee these extractions are all correct. We developed a web application that coordinates the ingenuity of humans in parallel to solve EHR information extraction problems that are difficult to automate.

Methods

We used medTurk to coordinate the answering of questions on what cancer a child was diagnosed with and what medical problems consistent with known late effects were observed. We analyzed an IRB approved EHR dataset comprised of 27,533 unstructured clinical notes of 243 pediatric cancer patients from Georgetown University's EMR system using medTurk in one week.

Results and Discussion

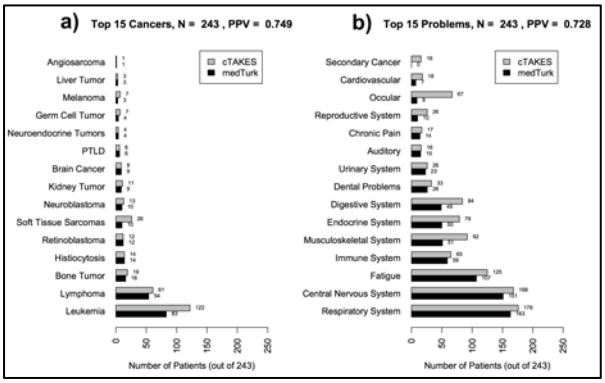


Figure 1. Comparison of results from uninspected cTAKES extractions and medTurk for a) Cancer diagnoses and b) problems observed in patients. The graph shows number of patients associated with that particular problem or cancer for each approach. The PPV value is calculated using medTurk as a gold standard.

In our particular use-case, if manual extractions were not performed via medTurk and cTAKES was relied on instead, then PPV (True Positives / [True Positives + False Positives]) would be reduced by 25% (see Figure 1). medTurk, although still in active development, is available at https://github.com/ICBI/medturk.

Identifying Candidate Ataxin-2 Inhibitors in High-Throughput Screening Data Using Molecular Descriptors

David E. Jones [1], Jingran Wen [1], Daniel R. Scoles [2], Thomas Dexheimer [3], Hongmao Sun [3], David Maloney [3], Anton Simeonov [3], Ajit Jadhav [3], Stefan M. Pulst [2] and Julio C. Facelli [1]

[1] Department of Biomedical Informatics and [2] Department of Neurology, University of Utah, Salt Lake City, UT and [3] National Center for Advancing Translational Sciences (NCATS), Rockville, MD

Summary: This presentation shows that machine learning and data mining can be used effectively to analyze data from high-throughput screening (HTS) for finding promising drugs for repurposing and new promising molecular entities. By using molecular descriptors of the chemical entities, the method used here is able to identify molecular motifs characteristic of candidate inhibitors for treating spinocerebellar ataxia type 2 (SCA2).

Introduction: SCA2 is a neurodegenerative disease with no known treatment. SCA2 is caused by CAG expansion in an encoded ATXN2 region resulting in polyglutamine expanded ataxin-2 protein with toxic gain of function. We hypothesize that lowering ATXN2 expression would be therapeutic for SCA2. HTS provides a powerful way to test the response of the expression of ATXN2 genes to chemical compounds. However efficient methods are needed to identify potential inhibitors of ATXN2 expression from screening data.

Methods: 1407 compounds in the NCGC pharmaceutical collection were screened with an HTS assay for response of ATXN2 gene expression and cytotoxicity. Compounds were clustered based on their screening response, chemical properties represented by 60 molecular descriptors and both types of variable using k-means clustering, with k=26 selected according to the literature (Mardia KV, Kent JT, and Bibby JM, Multivariate Analysis; in Probability and Mathematical Statistics, Academic Press, London 1979). Chemical property patterns in the clusters with high inhibition of ATXN2 expression and low inhibition of CMV-luc and biological control were analyzed. The compounds in these clusters could be considered as candidates for SCA2 treatment and their common chemical motifs may be used to design novel candidate inhibitors.

Results: The analysis of the results shows that the cluster with the most desirable properties contains 16 compounds (molecular descriptor only clustering); this cluster is also the most desirable when including the screening responses showing that the compounds in the cluster not only have consistent chemical properties, but also similar screening responses. The compounds in this cluster show statistical significance in exhibiting high inhibition of ATXN2-luc expression (average = -88.6 vs. -75.2 for the rest of the compounds) with minimal inhibition of CMV-luc expression (average = -43.5 vs. -60.4 for the rest of the compounds) and biological control expression (average = -2.3 vs. -10.0 for the rest of the compounds). The key molecular descriptors characterizing this cluster, which have been determined by a t-test analysis to be statistically significant, are: larger compounds (average mass of compounds in the cluster 624 Daltons vs. 387 Daltons in the rest of the compounds); more polarizable compounds (average molecular polarizability 63.8 $Å^3$ for the compounds in the cluster vs. 40.3 $Å^3$ in the rest of the compounds) and compounds less hydrophobic (partition coefficient, logP, for the compounds in the cluster of interest is 1.98 vs. 3.56 for the rest of the compounds). Further analysis of the structural features indicates that the compounds in the most desirable cluster typically have a core structure of large, fused aliphatic rings with attached carbonyl groups.

Discussion: Machine learning and data mining can be used effectively to analyze data from HTS for the repurposing of drugs. The method used here is able to clearly identify candidate inhibitors for treating spinocerebellar ataxia type 2. These candidate inhibitors have common molecular descriptors and molecular motifs that may be used to design novel inhibitors with similar molecular structures.

Acknowledgements: DEJ has been supported by NLM training grant 5T15LM007124, JW was partially supported by the Fay's Informatics Research Fellowship. The authors acknowledge NIH grants RC4NS073009, R21NS081182, and R01NS033123 (DRS and SMP), NCATS U54MH084681 (TD, HS, DM, AS, AJ) and NLM 5T15LM007124 and NCATS 1ULTR001067 (JCF).

A Risk Model for 30-Day Heart Failure Re-Admission using Electronic Medical Records

Uri Kartoun, Ph.D.[1,2], Vishesh Kumar, MD[1,2], Ari Brettman, MD[1,2], Sheng Yu, Ph.D.[3], Katherine Liao, MD, MPH[2,4], Elizabeth Karlson, MD[2,4], Ashwin Ananthakrishnan, MBBS, MPH[2,5], Zongqi Xia, MD, Ph.D.[2,5], Vivian Gainer, M.S.[6], Andrew Cagan, B.Sc.[6], Shawn Murphy, MD, Ph.D.[6], Susanne Churchill, Ph.D.[5], Isaac Kohane, MD, Ph.D.[2,5], Peter Szolovits, Ph.D.[7], Tianxi Cai, Sc.D.[3], Stanley Y. Shaw, MD, Ph.D.[1,2]

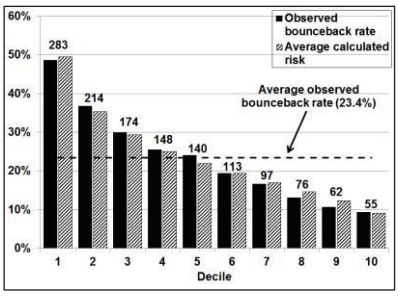
1. Center for Systems Biology, Massachusetts General Hospital (MGH), Boston, MA; 2. Harvard Medical School, Boston, MA; 3. Department of Biostatistics, Harvard School of Public Health, Boston, MA; 4. Division of Rheumatology, Brigham and Women's Hospital (BWH), Boston, MA.; 5. i2b2 National Center for Biomedical Computing, BWH, Boston, MA; 6. Research Computing, MGH, Boston, MA; 7. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Summary: We develop a risk score for re-admission following an index congestive heart failure (CHF) admission, using codified and narrative variables from 17 years of electronic medical records (EMR). We show a strong correlation between calculated and observed re-admission frequencies throughout the range of risk, in both diabetic and non-diabetic populations.

Introduction/Background: Identifying patients at high risk of re-admission after an index admission for CHF is of major academic, operational and financial interest. We hypothesized that an unbiased method of risk discovery drawn from EMRs may identify patients at high risk for re-admission and provide opportunities for intervention.

Methods: We predicted the likelihood of an index CHF admission being followed by a subsequent admission for any cause within 30 days of discharge, using data available at two time points within the index admission: 1) the first 24 hours ("early"), and 2) at the time of discharge ("discharge"). Our study data included 17 years of inpatient CHF admissions at two urban tertiary care hospitals between 1993 - 2010. We focus on two cohorts: 1) 65,099 type-2 diabetes (T2D) patients, with 5,825 index CHF admissions (with 23.4% 30-day re-admission rate), and 2) a *Non-Diabetic Cohort* of 43,220 patients (2,203 index CHF admissions and 22.4% 30-day re-admission rate). We extracted 293 EMR variables including demographics, laboratory values and slopes, billing codes, cardiac parameters extracted from narrative electrocardiogram and echocardiographic reports, and medical concepts extracted from physician narrative notes using natural language processing.

Results: Using logistic regression with the adaptive LASSO, we found a strong correlation between predicted and observed risk of re-admission throughout the range of calculated risk for the *Diabetic Cohort* ($r \geq 0.99$ for both the "early" and "discharge" models). Patients who had a re-admission within 30 days had a significantly higher predicted risk score vs. patients who were not re-admitted ("early": 28.6% vs. 21.8%; $p = 3.7 \cdot 10^{-66}$, "discharge": 29.4% vs. 21.5%; $p = 2.7 \cdot 10^{-77}$). Using a four-fold cross validation scheme yielded C-statistics of 0.65 and 0.67 for the "early" and "discharge" models, respectively. The "early" and "discharge" models had comparable accuracy in assigning patients to the highest and lowest deciles of re-admission risk. Significantly, the "discharge" model successfully re-classified a subset of patients of intermediate risk in the "early" model: calculated and observed re-admission rates for patients re-classified into the highest-risk decile in the "discharge" model were 45.0% and 43.6%, respectively. Applying an analogous approach to the *Non-Diabetic Cohort* yielded similar results.



Discussion: A generalizable method using unbiased variable selection and model building from EMR data can successfully identify patients at high or low risk of re-admission. "Early" data can identify high and low-risk groups; additional data generated during the admission and available at time of discharge can further re-classify additional individuals into high or low risk groups. This two-phase approach to risk estimation may facilitate intervention for high-risk patients earlier in the index hospital admission.

A Toolkit Enabling Efficient, Repeatable Phenotype Algorithm Development and Sharing

Jacqueline Kirby, MS[1]; Luke Rasmussen[2]; Peter Speltz[1]; Jyotishman Pathak, PhD[3]; Jennifer A. Pacheco[2]; Will Thompson[4], PhD; James Cowan[1], Susan Osgood[1], Lisa Bastarache[1], Peggy Peissig, PhD[5]; Sarah Stallings, PhD[1]; Paul Harris, PhD[1]; Melissa Basford, MBA[1], Josh Denny, MD MS[1]

[1]Vanderbilt University, Nashville, TN; [2]Northwestern University, Chicago, Illinois; [3]Mayo Clinic, Rochester, Minnesota; [4]NorthShore University Health System, Evanston, IL; [5]Marshfield Clinic Research Foundation, Marshfield, Wisconsin

Summary: Developing sharable, efficient, and accurate phenotype algorithms for electronic health records is time consuming and challenging. The Electronic Medical Records and GEnomics (eMERGE) Network has refined a workflow and suite of tools that assist in this process. Here we describe the general workflow, barriers, and tools for phenotype algorithm development.

Introduction: The eMERGE Network and others have proved that using EMR data can identify disease phenotypes with sufficient positive and negative predictive values for use in clinical and genetic research. We have found that accurate, sharable and quickly adaptable phenotype algorithms in varying clinical data repositories are both needed and difficult. Lessons learned in multisite collaboration for phenotype algorithm development include early assessment of feasibility for implementation, appropriate versioning, standardizing and reusing data elements for quality and efficiency, data quality and validation checks, and timing and methods for disseminating the results.

Results: The eMERGE Network has designed a Phenotyping Toolkit to facilitate and standardize workflow for all stages of developing and sharing phenotype algorithms. The Phenotype KnowledgeBase (PheKB.org) was functionally designed to enable such a workflow and has purposefully integrated tools and standards that guide the user in efficiently navigating each of these stages from early stage development to public sharing and reuse. To facilitate **Phenotype Algorithm Creation**, we have built a feasibility testing resource, the eMERGE RecordCounter (eRC), extended existing tools for refining and documenting algorithms in a computable format (PheWAS R-package, KNIME), and a data dictionary standardization tool (eleMAP). The eRC facilitates data exploration and testing of coarse phenotype assumptions using EMR-derived demographic data, ICD9 codes, and CPT codes linked to genotyped samples. The eRC plays a valuable role by enabling investigators to determine if an approach is feasible given the available data and its phenotypic characteristics. Users develop feasible phenotype algorithms and share on PheKB for iterative feedback and revision. Algorithms can be in varied formats, such as word documents or open-source data mining platform Konstanz Information Miner (KNIME.org) workflows. eleMAP is used to clearly define and standardize algorithms. The R PheWAS packages enable others to easily perform PheWAS. The **Phenotype Sharing and Validation** stage uses embedded PheKB tools to enable cross-site collaboration for algorithm development and validation. These include electronic discussion and update notifications for a given phenotype; the ability to post algorithm validation details and tools for validation including chart review forms and automated calculation of validation statistics such as specificity and sensitivity; and an integrated data dictionary and dataset validation tool used to ensure consistent formatting and coding of data sets to be shared. In addition, built-in privacy within PheKB enables algorithm owners to determine view and edit rights and access shared data. Once an algorithm has successfully been validated, the next stage is **External Implementation** at other collaborating sites who contribute study specific data. Site-specific implementation feasibility is accomplished by utilizing the eRC. Data validation and sharing tools built into PheKB encourage data standardization and early stage quality control to efficiently share data for the merging of study sets. Finally, **Dissemination** tools allow the sharing of results and work for reuse. Within PheKB one can publicly share algorithms as well as multiple implementation results with recorded PPV, NPV and site-specific notations. This allows reuse of the algorithms with confidence and baseline implementations comparisons. Other dissemination tools include a variant repository tool for exploring drug response implications of genetic variation (emergeSPHINX.org) and a catalog of phenome-wide association studies that analyze many phenotypes compared to a single genetic variant or other attribute (phewascatalog.org).

Conclusion: While barriers exist in creating and sharing reusable, standardized phenotypes, the framework and toolkit described here represent a proven model supporting 21 completed and 73 in-progress phenotypes across the eMERGE Network.

NLP-TAB: A system for integration and visualization of diverse biomedical NLP applications.

Benjamin C. Knoll[1], Genevieve B. Melton, MD[1], Hongfang Liu, PhD[2], Hua Xu, PhD[3], Serguei V.S. Pakhomov, PhD[1]
[2]Mayo Clinic, Rochester, MN;
[3]The University of Texas Health Science Center at Houston, Houston, TX

Introduction

Direct comparison and evaluation of biomedical Natural Language Processing (NLP) systems created by different research groups using disparate data type systems is a challenging problem. Currently, a large amount of work is required to convert these applications to a common type system—a task that requires in-depth knowledge of the implementation of each NLP application and presents a significant challenge to NLP system interoperability and rapid system development. We present NLP Type and Annotation Browser (NLP-TAB), an open-source system that facilitates exploration and analysis of NLP applications and their components without prior knowledge of their implementation. By storing and analyzing the results produced by each NLP application on one or more corpora using a type-agnostic data model, we allow users to discover which annotations best match their specific information retrieval tasks, as well as run comparisons between annotation types of separate applications. The ultimate goal of NLP-TAB is to facilitate the development and deployment of information extraction systems that make use of the results of multiple NLP applications developed using the Apache Unstructured Information Management (UIMA) platform (http://uima.apache.org/), maximizing their relative strengths and minimizing their weaknesses.

Background

Due to differences in the ways that NLP systems extract information from unstructured texts, direct one-to-one mapping between annotations produced by different NLP applications is not feasible without translating data types used by each system into a common format. As such, obtaining valid comparisons typically requires time-consuming conversions and/or modifications to various components of NLP applications and their type systems. NLP-TAB is designed to achieve a threefold purpose. First, it allows users to explore and evaluate disparate NLP applications and the annotations they create through several visualization and information retrieval techniques. Second, it allows combining the annotations produced by different NLP systems for subsequent information retrieval. Third, NLP-TAB enables the reuse and interoperability of components from different NLP pipelines through analysis and supervised or unsupervised creation of mappings between data types.

Methods

NLP-TAB consists of three components. The first is an UIMA analysis engine that receives annotation data from individual NLP applications, converts it to the type-agnostic format, and indexes it on an Elastic Search server. The second is a Java application that performs statistical analysis on the annotations indexed by the first component. This component uses the Elastic Search index to identify co-located annotations produced by NLP pipelines and computes the frequency of their occurrence and co-occurrence within a set of flexible boundaries. Based on frequency information, several measures of association are computed such as Matthews, Jaccard and F-score measures to show the strength of relatedness between annotations. The third component is a web application, which provides with visualization of the data created by the previous two components and the ability to navigate across NLP pipelines, their type systems and documents annotated with these pipelines.

Results

To our knowledge, NLP-TAB is the first system of its kind that requires no configuration and no prior knowledge of the NLP applications being compared or integrated. The primary components of NLP-TAB have complete functionality and are still undergoing active development. We have taken two systems, MetaMap and BioNLP, and have run them on a corpus of Genia documents and uploaded the results to NLP-TAB to provide a demo of functionality available here: http://athena.ahc.umn.edu/bionlptab.

Ability to leverage annotations from multiple NLP applications can potentially improve accuracy and reliability of information extraction from biomedical texts, particularly when the NLP applications produce complementary annotations. NLP-TAB is designed to measure the degree to which different NLP applications are complementary and enable combining their annotations for information retrieval.

Brain Injury Cognitive Screening (BICS): A simple, quick and effective application for screening mild traumatic brain injury

Jitendra Kumar MCA[1,2], Suresh Kumar, MD[1], Pooja Shah[1], and Ajay Jawahar, MD, MS[1]

1. Neuro-headache and TBI Rehabilitation Center, Shreveport, Louisiana, USA
2. Dayanand Saraswati University, Ajmer, Rajasthan, India

Objective: To present Brain Injury Cognitive Screening (BICS) as a mobile phone or tablet android application to quickly assess the cognitive impairment associated with mild traumatic brain injury.

Introduction: According to existing data, more than 1.5 million people experience a traumatic brain injury (TBI) each year in the United States and as many as 75 percent sustain a mild traumatic brain injury—or MTBI. These injuries may cause long-term or permanent impairments and disabilities. Many people with MTBI have difficulty returning to routine, daily activities and may be unable to return to work for many weeks or months. In addition to the human toll of these injuries, MTBI costs the nation nearly $17 billion each year. Each year in the United States 1.365 million are treated and released from an emergency department with the diagnosis of MTBI. MTBI can cause a wide range of functional changes affecting thinking, language, learning, emotions, behavior, and/or sensation. It can also cause epilepsy and increase the risk for conditions such as Alzheimer's disease, Parkinson's disease, and other brain disorders that become more prevalent with age. The Centers for Disease Control and Prevention estimates that at least 5.3 million Americans currently have a long-term or lifelong need for help to perform activities of daily living as a result of a MTBI.

Background: Unlike a major traumatic brain injury, concussion or mild TBI does not have overt clinical and/or radiological manifestations that make the diagnosis obvious. The radiological examination including the computerized tomography (CT) scan are usually normal and the clinical examination does not reveal any obvious abnormality. More often than not, it is the cognitive function of the brain that tends to get affected following a mild traumatic injury. The cognitive functions can be assessed in detail by several available test batteries but all of them require more than an hour and have to be administered and interpreted by qualified psychologists, who are usually not available in an emergent setting. This makes cognitive assessment of MTBI patients a challenging task in the setting of the accident site itself or even the Emergency Room. Montreal Cognitive Assessment (MoCA) is a quick screening tool that was devised by Nasreddine et al in 2005 to gauge cognitive impairment in patients with dementia. This test makes a quick assessment of 7 different functions, namely: visuo-spatial; naming; memory; attention; language; abstraction and orientation and allocates individual scores. Since then it has been validate as a reasonably reliable tool in stroke,

Parkinson's disease and a variety of neurological disorders affecting cognitive functioning. Our group recently examined the validity and relevance of MoCA in 130 patients of MTBI treated at out center. We concluded that 97% of the patients of MTBI demonstrated impairment of one of the four cognitive functions, namely: visuo-spatial orientation, language, delayed recall and attention span. We therefore attempted to develop a simple and user friendly android application that could quickly assess these four cognitive functions to enable first responders to assess and suspect the extent of brain injury after concussion or mild trauma in patients.

Methodology: Brain Injury cognitive Screening cell phone or tablet application has a simple interface for displaying the question and allowing the user to answer. It has audio or text interface to quickly assess the brain injury. It takes 1.5 minute to self assess or supervise the assessment of the brain injury. The application runs on all flavors of computers of relatively recent vintage: Linux, Mac OS, or Microsoft Windows. An Android device such as cell phone or tablet is useful (and of course the ultimate target for development), but is in fact not essential to getting started since the software contains virtual device emulators. The software has been written in American English language. It allots the scores of 0-12 to the patients based upon their responses to a few simple tasks. An impaired score of 9 or less should alert the assessors (first responders, EMT, ER physicians, athletic team coaches and sports physicians) to the possibility of cognitive impairment as a result of brain injury and cause them to seek detailed medical evaluation by physician with expertise in the area of TBI. Conclusions: Prevalence of Mild traumatic brain injury and its cognitive aftermaths has increased significantly in the past decade and has become the cause of increased social, economic and financial losses. MTBI is often missed as a diagnosis due to the absence of over clinical or radiological signs. A simple, quick and sensitive cell phone application like Brain Injury Cognitive Screening (BICS) will prove extremely effective in suspecting MTBI in relatively minor trauma setting and encouraging the assessing personnel to seek expert medical help for the victims.

Learning Objectives:

After attending this presentation the attendees should be able to:

1. Appreciate the increasing prevalence of Mild traumatic Brain Injury and the social and economic burden of the aftermaths of MTBI or concussion injuries.

2. Realize the need for a quick, effective and simple screening tool to indicate the suspicion of brain injury in even apparently minor accidents so the victims could be advised to seek expert help.

3. Assess the efficacy of Brain Injury Cognitive Screening (BICS) cell phone application in revealing the cognitive impairment associated with mild brain trauma in patients with concussion injury.

Constructing a UIMA-based High-performance Temporal Link Detection Pipeline with Rich Features

Ding-Cheng Li PhD[1], Guoxi Yan MS[2], Ravikumar Komandur Elayavilli, PhD[1], Majid Rastegar Mojarad, PhD[1], Yanpeng Li, PhD[1], Kavishwar B Wagholikar, PhD[1], Sungwan Sohn, PhD[1], Hongfang Liu, PhD[1]
Mayo Clinic, Rochester, MN US[1], The George Washington University, Washington, DC, 20052[2]

Abstract. *Temporal resolution for events and time expressions in clinical notes is crucial for an accurate summary of patient history, better medical treatment, and further clinical study. Discovery of a temporal relation aims at building a temporal link (TLINK) between events or between events and time expressions. TLINK detection in clinical natural language processing (NLP) improves clinical text mining and Clinical practice and research would benefit greatly from TLINK detection. In this work, we constructed a UIMA-based pipeline to detect TLINK among clinical notes. Rich features were extracted via this pipeline and liblinear classifier was employed to train the prediction model. The initial evaluation result on the 2012 I2B2 TLINK data sets reached 63% in F-score, which was close to the state-of-the art. In future work, we will explore more discriminative features so that a practical end-to-end system will be constructed.*

Background and Introduction. Similar to the general domain, the task of finding temporal relations between medical events in a clinical note is about finding whether an event is before, after, or overlap with another event. This process is usually called TLINK detection. Previous studies have explored diverse knowledge sources and linguistic information in inferring TLINK among events, including temporal adverbials, tense, aspects, rhetorical relations, pragmatic conventions, and background knowledge. Mani et al [1] employed machine-learning models to detect TLINK. They trained a MaxEnt model with features such as event class, aspect, modality, tense, and negation as well as event string and contextual features. Similar systems have been developed in medical TLINK detection system as well. Tang et al. [2] built a system composed of incremental classifiers, which yielded the state-of-art results in 2012 I2B2 TLINK Detection Challenge.

Methods. In essence, TLINK detection consists of two tasks: TLINK pair generations and TLINK assignments. Inspired by previous works, including our existing system [3], we designed a highly integrated UIMA-based TLINK detection system. Our system firstly links events and/or time expressions to form TLINK candidate pairs based on rules and next utilizes training data to train a TLINK prediction model to predict TLINK types on testing data. The beauty of this system lies in several points. Firstly, the candidate pair generations and feature generations are completed in a streamline fashion through aggregate UIMA analysis engines. Secondly, we build multiple classifiers for the TLINK detection, including event-admission classifier, event-discharge classifier, intra-sentential classifier and inter-sentential classifier via UIMA CasConsumer. The Liblinear model is employed to train classifiers and make predictions on the testing data efficiently. Thirdly, rich sources of features are extracted for each classifier. We use three sources for feature generation: 1) rule-based TLINK labels determined by the rules from the rule-based framework; 2) attributes extracted from the Clinical Narrative Temporal Relation Ontology—that is, concept definitions about time (eg, temporal instant, temporal interval) and relations between the time concepts (eg, prior to, before, earlier, after, until, subsequent, overlap, during, etc); 3) rich linguistic features. All of them are utilized to boost the performances. In particular, compared with our previous system, we strengthened our model with topic and distributional semantics features and TLINK labels obtained from previous classifiers.

Case Studies. For better comparisons, we also use 2012 I2B2 TLINK datasets for evaluations. The training set consists of 190 discharge summaries, and has been manually annotated for Events, TIMEX3s, and TLINKs. The test set includes 120 discharge summaries from the same sources. The annotations for TLINK events include relations within a single sentence/clause and beyond. The designing of our pipeline is partially inspired by the characteristics of the clinical notes. It is known that most i2b2 discharge summaries are comprised of two major sections including admission and discharge. The section information can be a strong clue to the order of the two events. For example, laboratory tests are generally performed before diagnosis. Such knowledge can be encoded as simple rules. Section information also makes it necessary to build separate classifiers for event-admission, event-discharge and others.

Results and Conclusion. The system successfully grouped events and time expressions into TLINK candidates and sequentially trained a high-performance prediction model. Our initial result reached 63% in F-score, close to the state-of-the-art. It is found that LVG, context features, section head, time-related features and topic features play prime roles. The contributions of the four classifiers are 19%, 21%, 18% and 5% respectively. In future work, we will continue to discover more discriminative features and do suitable post-processing.

Reference:

[1] I. Mani and J. Pustejovsky, "Temporal discourse models for narrative structure," in *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, 2004, pp. 57-64.

[2] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, "A hybrid system for temporal information extraction from clinical text," *Journal of the American Medical Informatics Association,* pp. amiajnl-2013-001635, 2013.

[3] S. Sohn, K. B. Wagholikar, D. Li, S. R. Jonnalagadda, C. Tao, R. K. Elayavilli*, et al.*, "Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification," *Journal of the American Medical Informatics Association,* 2013.

DNA Palindromes in Human Genome

Helen Li[1], Aman Gupta[2], BS, Madhavi K. Ganapathiraju, Ph.D[1,2]
[1]Department of Biomedical Informatics, University of Pittsburgh
[2]Language Technologies Institute, Carnegie Mellon University

Summary: *A palindrome in DNA is a sequence whose one half is complement of the other half but appearing in reverse order (for example: ATCC-GGAT). Palindromic sequences can influence DNA stability and are known to be associated with diseases. Here, we present all DNA palindromes in human genome that are equal to or longer than 8 bp. There are more than 39 thousand palindromes that are longer than 50 bp.*

Introduction and Background: Palindromic sequences are believed to be distributed throughout the human genome in varying lengths and locations. Palindromes are linked to genetic disorders, such as mental retardation, X-linked recessive diseases, and many physical abnormalities caused by mutagenesis. The prevalence and function of palindromes in the human genome is not fully understood. We present our findings on the distribution of palindromes in human genome, which reveals orders of magnitude of palindromes compared to a previous study.

Methods: We had previously developed Biological Language Modeling Toolkit (v. 2) that constructs suffix array, and longest common prefix array to efficiently identify all palindromic sequences in the human genome. We employed the toolkit on UCSC Genome Browser human genome Build 38 (GRCh38/hg38) and Build 19 (GRCh37/hg19), and analyzed all palindromes and near palindromes 8 bps and longer, allowing up to four mismatches. Evidence suggests that palindrome distribution is non-uniform in gene functional regions. We defined our functional regions: exons, introns, upstream as the region -2000 from the transcription start site, intergenic region as the noncoding area between genes, 5' UTR as the region between gene transcription start site and coding region start site, 3' UTR as gene coding region end and transcription start end, and promoter as the region -200 bp from transcription start site. We computed counts in each of these regions in each chromosome.

Results and Discussion: Analysis of the palindromic lengths revealed that palindromes of length 16 are the most frequent across all chromosomes (Figure 1A). The number of palindromes was proportional to the chromosome length as well as to gene density. We found very high palindrome counts in the intronic, upstream and intergenic regions, which may be linked to transcription and replication processes and formation of functionally relevant secondary structures in pre-mRNA. In total, we found 32,976,219 palindromes in human genome Build 38 (GRCh38/hg38), with 74.99% of all palindromes being AT rich. We also found 39,501 palindromes of length greater than 50 bp, which is significantly greater than other palindrome finders that only discovered 3,500 palindromes longer than 50 bp. These longer palindromes were also predominantly AT rich (Figure 1B). High AT content in palindromes, which create palindromic AT rich repeats (PATRR), are correlated to DNA instability, breakage, and chromosomal translocation (Kato et al. (2006)). Our results are available for view as a track on the UCSC Human Genome Browser human genome Build 38 (GRCh38/hg38). We are currently analyzing genome sequences from TCGA data to understand the role of palindromes in structural alterations in DNA, as there is evidence to show that long palindromes occur frequently in human cancer cell lines in medulloblastoma (Tanaka et al. (2006)).

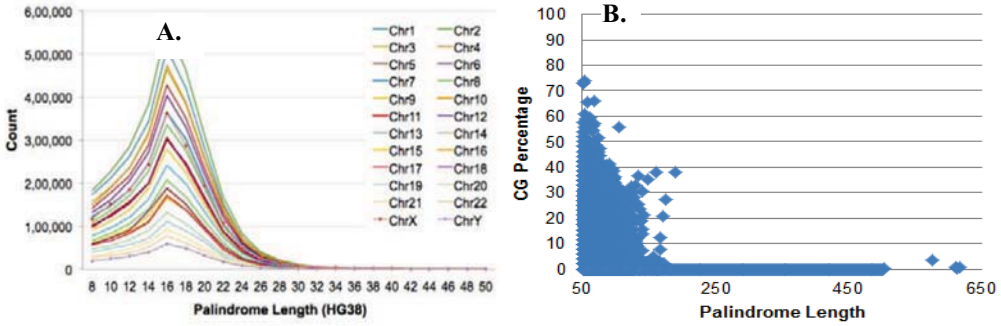


Figure 1 (A): Counts versus lengths of palindromes in each chromosome. **(B)** CG percentage versus

References

Guenthoer, J., *et al.* (2012) Assessment of palindromes as platforms for DNA amplification in breast cancer, *Genome research*, **22**, 232-245.
Tanaka, H., *et al.* (2006) Large DNA palindromes as a common form of structural chromosome aberrations in human cancers, *Human cell*, **19**, 17-23.
Kato T, Franconi CP, Sheridan MB, Hacker AM, Inagakai H, et al. (2014) Analysis of the t(3;8) of hereditary renal cell carcinoma: a palindrome-mediated translocation. Cancer Genet 207: 133-140

Identification of Similar Variables in the Database of Genotypes and Phenotypes (dbGaP)

Ko-Wei Lin, PhD, Son Doan, PhD, Alexander Hsieh, BS, Hyeoneui Kim, RN, MPH, PhD
Division of Biomedical Informatics, University of California San Diego, La Jolla, CA

Abstract

As a part of standardization effort of the phenotype variables in dbGaP, we developed a variable harmonizer that identifies the same variables with idiosyncratic descriptions. The performance varied by categories upon evaluation with 800 selected variables, and the average F-measure ranged from 0.62 in Laboratory Test to 1 in Demographics.

Introduction

We previously developed a new phenotype information retrieval system Phenotype Discoverer (PhenDisco, http://phendisco.ucsd.edu) for dbGaP (the database of Genotypes and Phenotypes http://www.ncbi.nlm.nih.gov/gap), a public databases of genome-wide association studies (GWAS). One of the core features of our system is phenotype standardization, which comprises of variable normalization and variable categorization modules[1]. This study reports on the variable harmonizer that we are adding to our standardization process. The main function of the variable harmonizer is to identify the variables that describe similar phenotype information in different ways. The variable harmonizer will facilitate the aggregation and harmonization of phenotypic data collected from different studies and presented in different formats in dbGaP.

Methods

Our standardization pipeline maps phenotype variable descriptions onto UMLS concepts and groups the variables into 16 categories[1]. We developed the variable harmonizer and evaluated with the four most frequently searched variable categories of Demographics, Medical History, Laboratory Test, and Medication. The variable categorizer identifies similar variables by comparing the topic concepts and the subject of information of the variables within a given variable category. We developed the harmonizer module using 88-convenience sample of variables in the 4 categories. We then evaluated it with randomly selected 800 variables falling into the 4 categories (i.e., 200 from each category). Three domain experts collaboratively reviewed the outputs of the harmonizer module. The performance of our variable harmonizer algorithms were evaluated using precision, recall, and F-measure.

Results and Discussion

Table 1. The average performance (s.d.: standard deviation)

	Precision (s.d.)	Recall (s.d.)	F-measure (s.d.)
Demographics	1 (0)	1 (0)	1 (0)
Medical History	0.86 (0.32)	0.89 (0.31)	0.87 (0.31)
Laboratory Test	0.67 (0.48)	0.60 (0.46)	0.62 (0.46)
Medication	0.92 (0.26)	0.88 (0.27)	0.89 (0.26)

The average performance metrics of the algorithms are presented in **Table 1.** The full data is available at https://idash-data.ucsd.edu/folder/3800. The results showed that the variable harmonizer could identify most of similar variables across different studies in dbGaP but the performance varied from category to category. A major contributing factor to lower performance is the failure or the inconsistency of topic concept identification step of our standardization pipeline. In particularly, several new laboratory test names could not be recognized by our pipeline, which resulted in the lower scores in the category of Laboratory Test. Future studies will include expanding our variable harmonizer to cover the other 12 categories in dbGaP and further testing the feasibility of applying PhenDisco standardization pipeline to other biomedical databases.

Acknowledgements This study was supported by the grant UH2HL108785 (NIH/NHLBI).

References

1. Doan S, Lin KW, Walker R, Farzaneh S, Alipanah N, Kim H. A Rule-based Natural Language Processing system in tagging and categorizing phenotype variables in NCBI's database of Genotypes and Phenotypes (dbGaP). 2013 AMIA Annual Symposium, pp.332, Washington DC, Nov. 16-20, 2013.

Improving Drug Safety Surveillance Using Chemical Systems Biology: Modular Assembly of Drug Safety Subnetworks

Tal Lorberbaum, BS [1,2,3], Mavra Nasir, MS [2,3], Michael J. Keiser, PhD [4,5,6], Santiago Vilar, PhD [2,3], George Hripcsak MD, MS [2,7], and Nicholas P. Tatonetti, PhD [2,3,7]

[1] Department of Physiology and Cellular Biophysics, [2] Department of Biomedical Informatics, [3] Departments of Systems Biology and Medicine, Columbia University, New York, NY; [4] Department of Pharmaceutical Chemistry, [5] Department of Bioengineering and Therapeutic Sciences, [6] Institute for Neurodegenerative Diseases, University of California San Francisco, San Francisco, CA; [7] Observational Health Data Science and Informatics, New York, NY

Abstract

To address long-standing limitations of traditional pharmacovigilance methods, we created an algorithm (Modular Assembly of Drug Safety Subnetworks) that integrates biological and chemical data to identify drugs with mechanistic connections to adverse events. By combining these chemical systems biology models with pharmacovigilance data we significantly improve drug safety monitoring for four clinically relevant adverse events.

Introduction

State-of-the-art pharmacovigilance algorithms continue to suffer from high false positive and false negative rates.[1] A recently published medication-wide association study (MWAS)[2] performed a self-controlled case series analysis using a gold standard of 150 drugs and four clinically relevant side effects (upper gastrointestinal bleeding (*GI*), acute liver failure (*LF*), acute myocardial infarction (*MI*), and acute kidney failure (*KF*)) yet was unable to eliminate many false positives and false negatives. Network analysis of molecular links between drugs and proteins that mediate adverse events offers a complementary approach.

Methods

We created an algorithm called the Modular Assembly of Drug Safety Subnetworks (MADSS) that utilizes protein-protein interaction (PPI) networks and machine learning to predict drug safety. For a given adverse event (AE), we modeled all human PPIs (STRING v9.1)[3] and annotated 8-35 "seed" proteins with a direct link to the target phenotype. We then adapted four network connectivity functions (mean first passage time, betweenness centrality, shared neighbors, and inverse shortest path) to score all other proteins in the PPI network on their connectivity to this seed set. Proteins receiving high connectivity scores constitute an "AE neighborhood" within the PPI network, and we hypothesized that drugs targeting proteins within this neighborhood are more likely to cause an AE. We assigned drugs in the MWAS gold standard the score of their most highly connected target. We then trained a random forest classifier using the drug scores from the four connectivity functions as features, resulting in four drug safety subnetwork (SubNet) models (one for each AE). In addition, we also grouped all AEs together to build a global model of adverse effects. Finally, we combined SubNet predictions with pharmacovigilance data (MWAS) using logistic regression (MWAS+SubNet) and validated using 10-fold cross-validation.

Results

We found that, individually, both MWAS ($\beta = 0.79 \pm 0.18$, $P = 1.05e\text{-}5$) and SubNet ($\beta = 4.34 \pm 0.58$, $P = 7.42e\text{-}14$) were significant predictors of adverse events. In addition, we found the combined model outperformed the univariate models ($\chi^2 = 75.9$, $P < 1 \times 10^{-15}$). For the combined model we found an AUROC of 0.85 compared to 0.81 and 0.69 for SubNet-alone and MWAS-alone, respectively (Figure 1A). In addition to outperforming overall, the combined model also outperformed for each adverse event individually with improvements in AUROC of 6.2% ($P = 0.10$), 33.9% ($P = 0.047$), 15.3% ($P = 0.01$), and 27.3% ($P = 0.007$) for GI, LF, MI, and KF, respectively (Figure 1B). At a false positive rate of 20%, recall improves from 42%, for MWAS alone, to 70% when drug safety statistics are combined with chemical systems biology data.

Discussion

Through the modular assembly of drug safety subnetworks, we demonstrate that chemical systems biology models can be successfully combined with current pharmacovigilance statistics across a range of etiologically diverse adverse events to improve predictive power. Post-marketing surveillance strategies should incorporate chemical systems biology models to enrich for potentially dangerous candidate drugs for follow-up study, ultimately reducing the health and economic impact of adverse events.

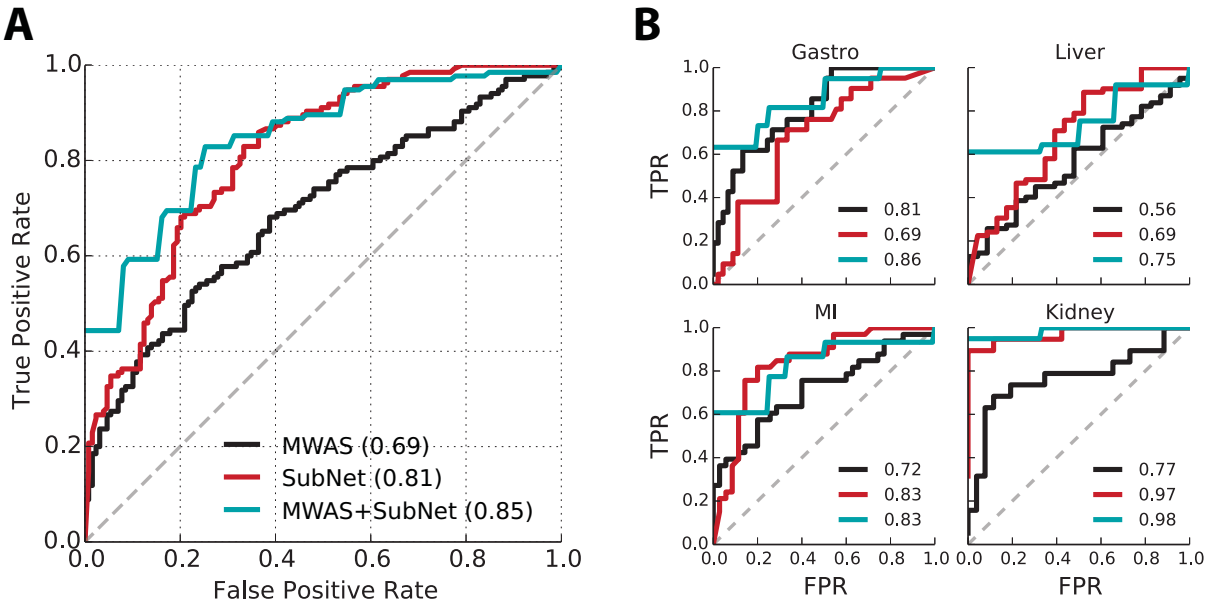


Figure 1. Chemical systems biology data significantly improve drug safety predictions. (**A**) ROC curve showing performance of pharmacovigilance statistics (MWAS) alone, chemical systems biology (SubNet) alone, and MWAS+SubNet for four adverse events (AEs) combined. AUROC is indicated in parentheses. (**B**) ROC curves demonstrating performance for individual AEs: gastrointestinal bleeding (Gastro), acute liver failure (Liver), acute myocardial infarction (MI), and acute kidney failure (Kidney).

References

1. Ryan, P. B. et al. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Statist. Med.* 31, 4401–4415 (2012).
2. Ryan, P. B., Madigan, D., Stang, P. E., Schuemie, M. J. & Hripcsak, G. Medication-Wide Association Studies. *CPT: Pharmacomet. Syst. Pharmacol.* 2, e76 (2013).
3. Franceschini, A. et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* 41, D808–D815 (2012).

Discovering Synergistic Multi-Drug Combinations for Breast Cancer

Yen S. Low, Alison W. Kurian, George W. Sledge, Jr., Tina Seto, Hastie Trevor, Nigam H. Shah

Abstract
Electronic health records (EHR) reflecting real-world treatment patterns can enable the discovery of new drug interactions. While the majority of drug interaction detection efforts focused on adverse interactions, we searched for beneficial interactions arising from synergistic multi-drug combinations associated with improved breast cancer survival. This approach allows us to optimize the use of existing drugs through appropriate drug combinations for optimal breast cancer outcomes.

Introduction
Large EHR databases reflecting real-world treatment patterns can enable the discovery of adverse drug effects, drug-drug interactions as well as unexpected beneficial drug effects. One such example is the purported anti-cancer benefit of the anti-diabetic drug metformin (1). Besides uncovering new indications for single drugs, EHR may harbor synergistic drug combinations whose combined therapeutic effect is better than that of individual drugs. In this study, we attempt to discover synergistic multi-drug combinations using an integrated breast cancer data resource (Oncoshare) which combines demographic and long-term outcomes from the California Cancer Registry with EHR from Stanford Hospital and Palo Alto Medical Foundation healthcare system.

Methods
Of the 9,956 patients followed for at least 5 years, we extracted 1,531 demographic, tumor, comorbidities and treatment variables including 1,455 drugs which were grouped into 100 drug classes according to the Anatomical Therapeutic Chemical system. Rare variables present in less than 1% of the patients were dropped. Pairs of drugs taken concomitantly were tested for synergistic and antagonistic effects on 5-year mortality using a lasso logistic regression with two-way interaction effects.

The model used an overlapped group lasso (2) such that main effects were selected along with the interaction effect. Variables selected by the lasso model were refitted to a final logistic regression and validated on a 10% hold-out set. Drug pairs with significant positive interaction effects were identified as synergistic drug pairs. Interacting drug pairs were visualized using networks. Community detection determined combinations of drugs that interacted together to moderate one another's effect on 5-year mortality.

Results & Discussion
Our lasso logistic regression model (93% AUC on hold-out set) identified protective (e.g. low stages, high income) and risk factors (e.g. triple negative breast cancer, comorbidities) consistent with prior knowledge. Interaction effects were more common among drug classes than individual drugs, possibly due to the former's higher prevalence. Beneficial interaction effects were detected: between higher stages and carboplatin or zoledronate; between anti-metabolites and anti-hemorrhagics, anti-hypertensives, or piperacillin; between anti-neoplastics and psycholeptics. One limitation of our linear model formulation is that we may have falsely selected correlated pairs of variables as pairs with interaction effects. Nevertheless, the model generated a set of potentially synergistic drug combinations for further evaluation by biological plausibility, epidemiological studies and clinical trials.

References
1. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. 2014 Jul 22;1–10.
2. Lim M, Hastie T. Learning interactions through hierarchical group-lasso regularization. 2013 Aug 12;1–35.

The AEExplorer, Its Developments and Applications in Clinical Trials, Especially for Investigating the Effects of Concomitant Proton Pump Inhibitors (PPIs) in Osteoporosis Clinical Trials

Luo, Zhongjun[1]; Yang, Jing[1]; Zuo, Zheng[1]; Xu, Nancy[4]; Whitaker, Marcea[5]; Tiwari, Ram[2]; Huang, Lan[2]; Luo, Zongwei[7]; Cooper, Charles[8]; Tong, Weida[6]; Rosario, Lilliam[1]; Navarro Almario, Eileen E [3]; Buckman-Garner, ShaAvhrée[3]; Office of Computational Science[1] and Office of Biostatistics[2] of Office of Translational Sciences[3] of Center for Drug Evaluation and Research, Division of Cardiovascular and Renal Products[4] and Division of Bone, Reproductive and Urologic Products[5] of Office of New Drugs of Center for Drug Evaluation and Research, Division of Bioinformatics and Biostatistics[6] of National Center for Toxicological Research, Food and Drug Administration (FDA). Department of Electrical and Electronics Engineer, South University of Science and Technology of China[7]. Infection Prevention and Management of Becton, Dickinson and Company[8].

Summary: The AEExplorer is a tool suite for the analysis and visualization of clinical trial study data, especially adverse events (AEs). The tool suite allows a user to select single or multiple clinical trials and any combination of AEs of interest. It displays in graphs the temporal relationships among the exposures of trial drugs, concomitant drugs, and the duration of AEs, regardless of trial size. It allows for conduction of basic statistical analyses such as relative risks (RRs) and associated confidence intervals (CIs) and more advanced analyses such as Likelihood Ratio Test (LRT).

Introduction and Background: We developed the AEExplorer to assess the effect of concomitant PPI use in the development of atypical fractures and other adverse events in osteoporosis clinical trials. Our first testing and use of the tool was prompted by the reports of paradoxical atypical femoral fractures and other osteoporosis-related adverse events in patients receiving drugs that treat osteoporosis, including bisphosphonates. The most frequent dose-limiting adverse effects of bisphosphonate use in the treatment of osteoporosis are nausea and gastrointestinal intolerance, which are often treated with PPIs [NIH Publication No. 09–4549].

Methods, Results, and Discussion: A database was populated with 13 clinical trials of osteoporosis drugs in standardized Study Data Tabulation Model (SDTM) format. Datasets were corrected for non-SDTM variable names, missing required values, missing required SDTM variables, and incorrect variable types. The clinical trials included were placebo-controlled with at least 3 years of total study drug exposure. The AEExplorer was developed to analyze the data and designed to allow: 1) selection of one or more trials and the AEs of interest; 2) graphic view of the temporal relationships among the exposures of trial drugs, concomitant drugs, and the duration of AEs, regardless of trial size; and 3) conduction of basic statistical analysis such as relative risks (RRs) and associated confidence intervals (CIs) that are conveniently displayed through tables and graphs, without running against SAS or any other statistical tool. The tool suite also runs a novel JAVA version of LRT (Lan Huang et al JASA 2014) that controls both the Type I error and false discovery rates while retaining good power and sensitivity for identifying signals. The LRT is useful especially for multiple trial analyses, as those in this study. The AEExplorer can be extended to study events other than adverse events, with or without concomitant drug uses, for a variety of clinical trial types. The only requirement for successful use of the tool suite is that the study data are loaded into the database, which in turn is accessible by the tool suite.

Conclusion: The AEExplorer is a useful tool suite for clinical trial study data analysis and visualization, especially for AE analysis. It allows analysis related not only to the trial drug but also to the concomitant drug. It displays temporal relationships among the exposures of trial drugs, concomitant drugs, and the duration of AEs, regardless of trial size. The tool suite augments the visualization capabilities with basic and advanced LRT statistical analysis.

The project is developed under a grant from the Office of Women's Health, Office of the Commissioner, FDA.

Druggability Profiling of Protein Biomarkers

Subramani Mani MBBS PhD[1], Daniel Cannon MS[1], Robin Ohls MD[1], Douglas Perkins PhD[1], Tudor Oprea MD PhD[1], Stephen Mathias PhD[1], Karri Ballard PhD[2], Oleg Ursu PhD[1], Cristian Bologa PhD[1]

[1]University of New Mexico Health Sciences Center, Albuquerque, NM, 87131

[2]Myriad RBM, Austin, TX 78759

Developing automated and interactive methods for building a mechanistic and potentially causal model of ranked biomarkers followed by incorporation into a contextual framework in disease-linked biochemical pathways can be used for potential drug-target evaluation and for proposing new drug targets. We demonstrate the potential of this approach using ranked biomarkers in the domain of neonatal sepsis obtained by enrolling 45 infants with late onset neonatal sepsis and 88 control infants.

Introduction and background: A set of predictive, diagnostic or prognostic biomarkers relevant to a specific clinical condition or disease can be identified using a focused literature search or obtained from research studies designed specifically for biomarker discovery. Since the proteome is downstream relative to the genome and transcriptome, assaying proteins holds considerable promise for discovering disease-linked biomarkers. It is clear that detection of multiparameter protein biomarkers from blood for prompt and early detection of systemic infections or cancers can lead to early treatment of the condition, resulting in more desirable outcomes. Though researchers have recognized the potential of multiple biomarker measurements for rationalizing the discovery of suitable drug targets(1), there is limited work on mechanistic modeling of the biomarkers and in identifying the set of druggable biomarkers(2-4) [that is, compounds/drugs that can modulate the protein(s) and/or receptor(s)]. Although biomarkers can play an effective role in early detection, diagnostic evaluation and for assessment of prognosis, a mechanistic understanding is required if biomarkers are to be targeted for druggability and eventual use of the biomarker(s) for therapeutics. The challenge is to identify the relevant subset of potential druggable targets that are represented in disease-linked proteins.

Methods: We performed a focused proteomic assay of 90 potential biomarkers suspected to play a role in infection and/or inflammation using serum samples collected from 45 cases of neonatal sepsis (culture positive) and 88 controls (culture negative) that we enrolled over a five year period from 2007 to 2012. The quantitative proteomic assay was performed by RBM using a customized implementation of the Luminex xMAP technology, a microsphere-based multiplexed immunoassay platform. Ranking of predictive biomarkers was performed using the Random Forest variable importance method to score and rank each variable. Using (1) pathway analytics and (2) the state-of-the-art drug database consisting of detailed information on approved and discontinued drugs worldwide (DRUGSDB)(5) developed by our group, we identified biomarkers that can serve as potential therapeutic targets for further evaluation and drug development. We have established a workflow to extract, merge and analyze pathways extracted from the "KEGG: Kyoto Encyclopedia of Genes and Genome" database(6, 7) that are relevant to biomarkers of interest. The method also involves the use of "R Bioconductor - KEGGgraph"(8-10) and “igraph”(11) (open source) statistical library. We also performed a target tissue localization (TTL) analysis of the top sepsis biomarkers using the Human Proteome Map (HPM)(12).

Results: Fifteen biomarkers were identified that predicted infection with high accuracy. We also identified five top ranked pathways extracted from the KEGG database that incorporate one or more of the top ranked sepsis biomarkers—Cytokine-cytokine receptor interaction, TNF signaling, Chagas disease, NOD-like receptor signaling and Cytosolic DNA-sensing pathways. Most of the top ranked biomarkers are cytokines and of the top selected pathways in sepsis the NOD-like receptor signaling pathway includes three biomarkers selected using the methods described in the Methods section. Some of the protein targets present in this pathway are modulated by approved and investigational drugs depicted in Figure 1. Figure 2 provides the relative expression of some of the top ranked sepsis biomarkers in 30 clinically defined healthy tissues (17 adult tissues, 6 primary hematopoietic cells and 7 fetal tissues). It also indicates the lack of expression for five of them in healthy tissues.

Discussion and Conclusion: Biomarker profiling for potential drug target identification using KEGG pathway analytics and the DRUGSDB database has the potential to identify drug-target interactions related to the biomarkers of a specific disease or clinical condition. The six drug-target interactions extracted from KEGG and the single drug-target interaction identified using DRUGSDB need further evaluation with respect to altering the pathology and clinical course of neonatal sepsis. The tissue-specific expression data shown in Figure 2 can be used to prioritize drug targets in terms of tissue relevance for many clinical conditions. This study opens up the possibility of biomarker druggability profiling using informatics methods and databases to propose new drug targets and for a systematic identification of drug-target interactions.

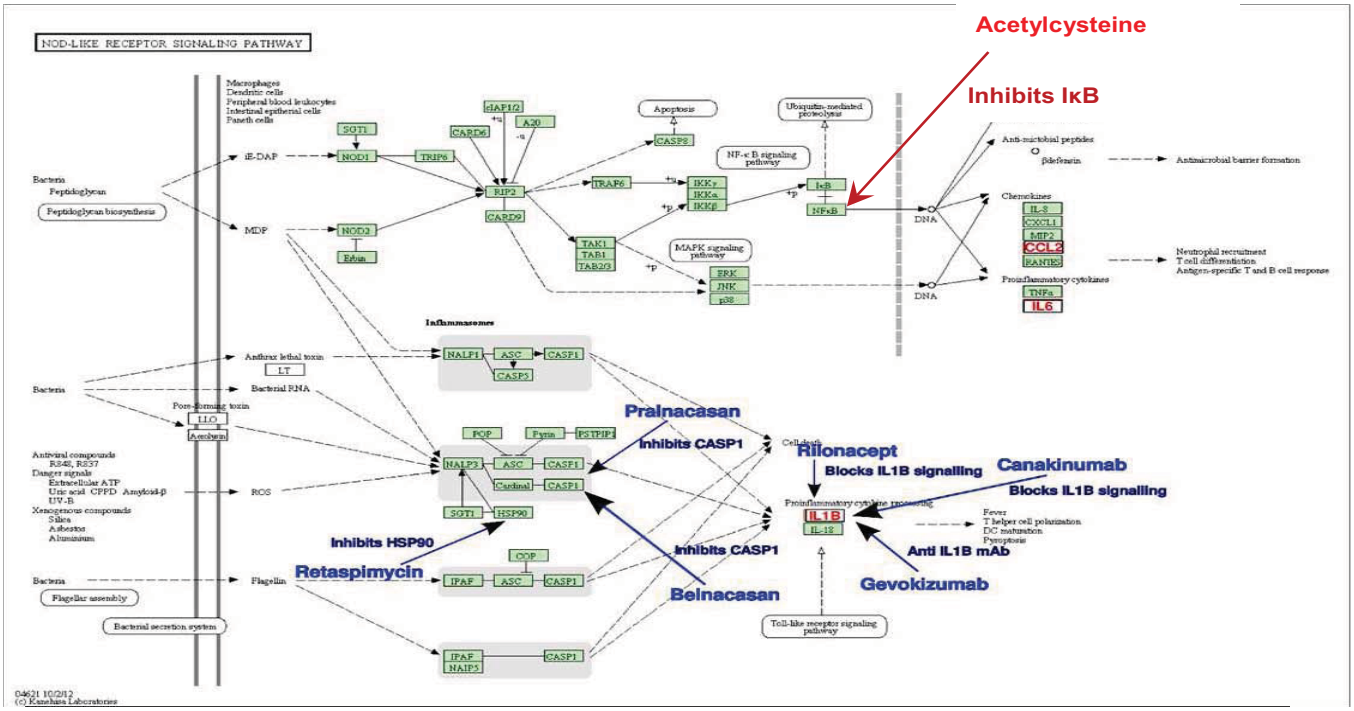


Figure 1: The NOD-like receptor signaling pathway selected from the top sepsis pathways. Blue: drug-target interactions directly extracted from KEGG. Red: drug-target interaction extracted from DrugBank.

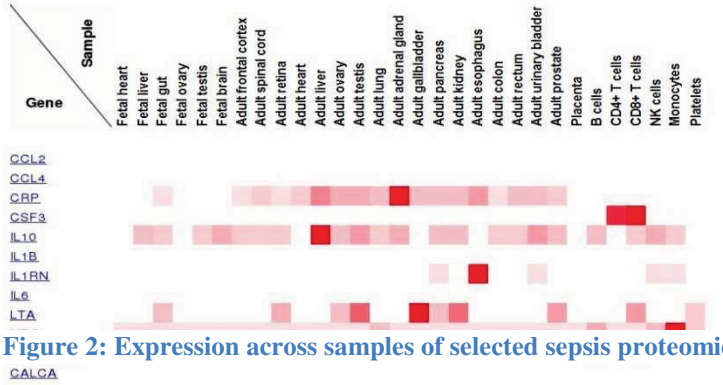


Figure 2: Expression across samples of selected sepsis proteomic

1.Hood L, Heath JR, Phelps ME, Lin B. Systems biology and new technologies enable predictive and preventative medicine. Science. 2004;306(5696):640-3.

2.Danhof M, Alvan G, Dahl SG, Kuhlmann J, Paintaud G. Mechanism-based pharmacokinetic–pharmacodynamic modeling—a new classification of biomarkers. Pharmaceutical research. 2005;22(9):1432-7.

3.Mayr M, Zhang J, Greene AS, Gutterman D, Perloff J, Ping P. Proteomics-based Development of Biomarkers in Cardiovascular Disease Mechanistic, Clinical, and Therapeutic Insights. Molecular & Cellular Proteomics. 2006;5(10):1853-64.

4.Shin D, Arthur G, Popescu M, Korkin D, Shyu C-R. Uncovering influence links in molecular knowledge networks to streamline personalized medicine. Journal of biomedical informatics. 2014.

5.Oprea TI, Nielsen SK, Ursu O, Yang JJ, Taboureau O, Mathias SL, et al. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. Molecular informatics. 2011;30(2-3):100-11.

6.Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic acids research. 2014;42(D1):D199-D205.

7.Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research. 2012;40(D1):D109-D14.

8.Team RC. R: a language and environment for statistical computing. 2013. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0; 2013.

9.Zhang JD, Wiemann S. KEGGgraph: a graph approach to KEGG PATHWAY in R and bioconductor. Bioinformatics. 2009;25(11):1470-1.

10.Zhang JD. KEGGgraph: Application ExamplesR package version 1.20.0. 2013.

11.Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695(5).

12.Kim M-S, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509(7502):575-81.

Informing the Hunt for Rare Disease Variants: The International Mouse Phenotyping Consortium

Terrence F Meehan[1], on behalf of the IMPC and the Monarch Initiative
1= EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

SUMMARY:
The **International Mouse Phenotyping Consortium (IMPC)** is characterizing 1000's of new knockout mouse strains for genes with little or no known biological function. Using algorithms developed by the Monarch Initiative, candidate causal genes for rare diseases are predicted based on overlapping phenotypes with IMPC strains.

Background:
The **International Mouse Phenotyping Consortium** is building a comprehensive functional catalogue of a mammalian genome by generating and broadly phenotyping a knockout mouse strain for every protein-coding gene. With phenotypes currently associated to over 1,375 genes, the IMPC resource can be used to help identify causative variants in human genetic diseases. The **Monarch Initiative** is using IMPC data as part of its mission to integrate, align, and re-distribute cross-species gene, genotype, variant, disease, and phenotype knowledge.

Methods:
To determine the degree of phenotype overlap between IMPC mouse strains and human genetic diseases, a Phenotype Similarity Score (PSS) was used as previously described (1). An automated pipeline generates PSS scores between IMPC strains and genetic diseases represented within Online Mendelian Inheritance in Man (OMIM), ORPHANET, and DECIPHER resources. PSS are indexed, categorized and integrated with mouse-human orthology mappings as supplied by ENSEMBL.

Results:
As of January 2015, phenotype data from 1375 strains have been analysed and PSS scores determined. A significant degree of phenotype overlap (PSS score > 50) is observed in 1233 IMPC mouse strains with over 3,000 diseases (3203 OMIM, 455 ORPHANET, 8 DECIPHER). Of particular interest are OMIM described diseases where a linkage locus has been determined but no causative variant identified. 93 IMPC strains have knockout genes that are syntenic to linkage loci associated with 75 OMIM diseases and include possible candidates for diseases such as "Arrhythmogenic Right Ventricular Dysplasia, Familial, 3" and "Pseudohypoaldosteronism, Type IIa". Results are displayed in an intuitive manner on the IMPC portal on dedicated disease and gene pages.

Discussion:
We demonstrate here that high-throughput phenotyping of a model organism can provide insight into determining human disease-causing variants. Resources such as the Exomiser tool developed by the Monarch Inititiave are extending this concept by using IMPC data to prioritize the dozens of candidates generated by exome analysis of rare disease patients. With another 3,500 mouse strains currently being phenotyped, the IMPC is becoming a rich resource for those wanting to better understand and test disease mechanisms.

1. Smedley,D., et al (2013) PhenoDigm: analyzing curated annotations to associate animal models with human diseases. *Database J. Biol. Databases Curation*, **2013** (online)

Weighting of Sensitivity and Specificity Estimates to Obtain Unbiased Performance Estimates for Classifiers using Stratified Test Sets

Douglas Morrison, MS[1], Tina-Hernandez-Boussard, PhD[1]
[1]Stanford University School of Medicine, Stanford, CA

Abstract

When test data for classifiers are collected using stratified sampling, performance metric estimates must be weighted using the population distribution of the stratifying variables, in order to avoid bias.

Introduction

When testing a classifier for a rare outcome whose gold standard value is expensive to determine, collecting a test dataset using simple random sampling would be an inefficient way to collect enough events to precisely estimate the classifier's sensitivity. Instead, we might stratify the test data sample to which we apply gold standard verification; for example, we might sample 100 records that the classifier labels as positive and 100 labeled negative. If so, in order to be unbiased, sensitivity and specificity estimates must be weighted using the population distribution of the stratifying variables. We use a simulation to illustrate this result.

Methods

We simulated a population of 10,000 observations with a 10% event rate, and simulated classifier results for this population with sensitivity = specificity = 80%. We collected a stratified sample of 100 subjects classified positive and 100 classified negative. We then calculated sensitivity and specificity using two methods:

(1) unweighted sample fractions:

$$sensitivity = \frac{\#\ true\ positives\ in\ sample}{\#\ true\ positives\ in\ sample\ +\#\ false\ negatives\ in\ sample}$$

$$specificity = \frac{\#\ true\ negatives\ in\ sample}{\#\ true\ negatives\ in\ sample\ +\#\ false\ positives\ in\ sample}$$

(2) inverse-sampling-probability-weighted estimates[1]:

$$w_P = \frac{\#\ in\ population\ classified\ as\ positive}{\#\ in\ sample\ classified\ as\ positive} = \frac{\#\ in\ population\ classified\ as\ positive}{100}$$

$$w_N = \frac{\#\ in\ population\ classified\ as\ negative}{\#\ in\ sample\ classified\ as\ negative} = \frac{\#\ in\ population\ classified\ as\ negative}{100}$$

$$sensitivity = \frac{\#\ true\ positives\ in\ sample * w_P}{(\#\ true\ positives\ in\ sample * w_P) + (\#\ false\ negatives\ in\ sample * w_N)}$$

$$specificity = \frac{\#\ true\ negatives\ in\ sample * w_N}{(\#\ true\ negatives\ in\ sample *\ w_N) + (\#\ false\ positives\ in\ sample *\ w_P)}$$

We repeated the simulation 1,000 times to estimate average performance for each metric and method.

Results

The unweighted sensitivity and specificity estimates had means ± standard deviations of 92±5% and 58±2%, respectively, while the weighted estimates averaged 81±10% and 80±1%, respectively. Paired t-test comparisons of the weighted and unweighted estimates had p-values < 0.0001.

Discussion

Classifiers are used in an ever-increasing range of important applications, and unbiased performance measurement is essential to ensuring that these tools function in practice as reliably as we expect them to. Our results show that stratified test data samples should be analyzed using statistically appropriate complex sampling methods.

Clinical Diagnoses Prediction with Temporal Data Analytics

Robert Moskovitch, PhD, Colin Walsh, MD, Hyunmi Choi, MD, George Hripcsak, MD, MS, Nicholas Tatonetti, PhD
Department of Biomedical Informatics, Columbia University, New York, USA
rm3198,cgw2106,hc323,gh13,npt2105@columbia.edu

ABSTRACT

The increasing availability of temporal data from Electronic Health Records (EHR) provides exceptional opportunities for the prediction of clinical outcome events. However, the nature of EHR data is sparse, incomplete and heterogeneous. We propose to transform the data into symbolic time intervals series, and then discover frequent Time Intervals Related Patterns (TIRPs) to predict the outcome events. In this study we rigorously focused on fifty frequent diagnoses, having very encouraging results.

INTRODUCTION

The increasing availability of time-stamped electronic health records (EHR) enables researchers to perform classification and prediction tasks that leverage the EHR's temporal nature -- one of the most challenging research topics in biomedicine[1]. This challenge arises since EHR data are sparse, incomplete, and stored in heterogeneous formats[1,2,3]. Therefore, for purposes of prediction, we propose to transform the time point series into symbolic time interval series, through a process known as knowledge based temporal abstraction[1,2], providing a uniform format of various temporal variables. Then, it is possible to discover frequent Time Intervals Related Patterns (TIRPs)[2,3,4]. The use of TIRPs as features for the classification of multivariate temporal data was proposed in[2,3,4]. In our study, the discovered TIRPs are used as features, to predict clinical diagnoses, which we introduce here and demonstrate on EHR data. Prediction and forecasting of diagnoses or visit utilization are important topics in the biomedical domain; these tasks span multiple clinical disciplines in medicine.

METHODS and RESULTS

We introduce Maitreya, a framework for Outcome Events Prediction, based on Time Intervals Related Patterns (TIRPs). Its main components are: the KarmaLego[2] and SingleKarmaLego[3] algorithms, used for frequent TIRPs discovery and detection, respectively. Maitreya offers two TIRP metrics in addition to the default Binary: horizontal support (the number of instances of a TIRP detected at a patient) and mean duration (the average duration of these instances), which will be used to represent the features in the prediction task. Our methodology consists of discovering TIRPs only in the set of patients having the outcome, which we call the Outcome Cohort. A Control set of patients is defined that are similar to the Cohort in their concepts data, which is used to induce a binary classifier and evaluate the method, as illustrated in figure 2.

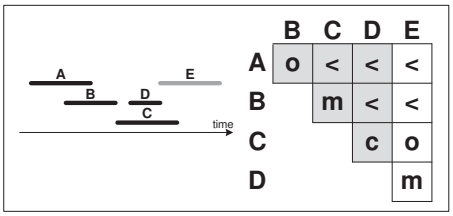


Figure 1. An example of a Time-Interval Related Pattern (TIRP), represented by a sequence of five symbolic time intervals and all of their pair-wise temporal relations.

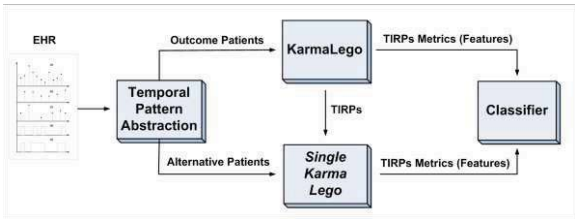


Figure 2. The EHR data is abstracted into symbolic time intervals and TIRPs are discovered. The patients represented by their TIRPs used to induce a classifier and evaluate their performance.

In order to evaluate our Outcome Events Prediction method we used the data from the Columbia University Medical Center New York Presbyterian Hospital (CUMC-NYP), containing approximately 30 million diagnosis billing codes, 20 million prescription orders, 9 million procedures, and 500 million laboratory results. We used only coded data for this analysis, including drug exposures, conditions (billing codes), and procedures. Medical concepts are then transformed into symbolic time intervals (called "eras"). We created datasets of thousands of patients having a diagnosis and their data based on a year, a month prior to the diagnosis, and control patients that are similar to the cohort. Our results are very encouraging, for example, in procedures we can predict hemodialysis (AUC=0.88) and blood transfusion (AUC=0.88).

REFERENCES

[1] G. Hripcsak, D. Albers, Next-Generation Phenotyping of Electronic Health Records, Journal of American Medical Informatics Association, 20: 117-121, 2013.

[2] R. Moskovitch, Y. Shahar, Fast Time Intervals Mining Using Transitivity of Temporal Relations, Knowledge and Information Systems, DOI 10.1007/s10115-013-0707-x, In Press, 2013.

[3] R. Moskovitch, Y. Shahar, Classification of Multivariate Time Series via Temporal Abstraction and Time Intervals Mining, Knowledge and Information Systems, In Press, 2014.

[4] I. Batal, D. Fradkin, J. Harrison, F. Moerchen, and M. Hauskrecht, Mining Recent Temporal Patterns for Event Detection in Multivariate Time Series Data, Proceedings of Knowledge Discovery and Data Mining (KDD), Beijing, China, 2012b.

Filtering Negative Reports for a Comparative Effectiveness Study of Stroke

Danielle L. Mowery, PhD[1], Brett R. South, PhD[1] Erin Madden, MPH[2],
Salomeh Keyhani, MD, MPH[2], Brian E. Chapman, PhD[1], Wendy W. Chapman, PhD[1]
[1]University of Utah, Salt Lake City Veteran Affairs, UT; [2]San Francisco Veteran Affairs, CA

Abstract: *We present a pilot study to develop an NLP pipeline that reduces chart review by filtering non-carotid images and negative carotid stenosis cases using machine learning and natural language processing.*

Introduction: 795,000 people suffer strokes each year for which about 12% of strokes are related to carotid artery stenosis. Creating patient cohorts with carotid artery stenosis is challenging because most carotid images are negative for stenosis. Manual chart abstraction is expensive and tedious limiting cohort creation for epidemiologic studies. Machine learning and natural language processing can help reduce manual review efforts by automatically filtering non-carotid images based on lexical, syntactic, and other features of clinical documents and negative reports for carotid stenosis based on anatomical location, negation, and severity of the finding description. For this study, we trained and evaluated a ML pipeline, *Automated Retrieval Console (ARC)*[1] and an NLP algorithm, *pyConText*[2,3].

Methods: We randomly selected 100 VA radiology images (RAD) and 100 VA progress notes (TIU) related to a specific date range an image was performed. For each report, 4 research associates independently abstracted (1) whether the report was a carotid image or not and (2) whether it contained a positive mention for significant carotid stenosis (severity > 50%). For each report type, we randomly sampled 75% of reports to train and 25% of reports to test two algorithms: task (1) conditional random field (ARC) to detect whether the report was a carotid image or not, and task (2) ARC and pyConText to assert whether the report contains a positive mention or not for carotid stenosis. We evaluated these system (S) performances against reference standard (RS) abstractions using sensitivity aka recall (probability that S+|RS+), positive predictive value (PPV) aka precision (probability that RS+|S+), specificity aka true negative rate (probability that S-|RS-), and negative predictive value (NPV) (probability that RS-|S-).

Results: 20% of train and 18% of test of RAD reports were not carotid images; in contrast to, 75% of train and 76% of TIU reports. 87% of train and 84% of test of RAD reports did not contain a significant carotid stenosis finding; similar to, 83% of train and 88% of TIU reports. For identifying negative carotid image reports on test, ARC achieved 100% specificity/75% NPV for RAD; 33% specificity/50% NPV for TIU reports. For identifying negative carotid stenosis reports on test, ARC achieved 91% specificity/91% NPV for RAD and 100% specificity/84% NPV for TIU. pyConText achieved 82% specificity/100% NPV for RAD and 90% specificity/100% NPV on TIU.

Table 1. Performances for each approach for tasks (1) and (2).

Carotid image or not	**Sensitivity**		**PPV**		**Specificity**		**NPV**	
ARC	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
RAD	94	95	89	100		100		75
TIU	98	95	95	91		33		50
Stenosis or not	**Sensitivity**		**PPV**		**Specificity**		**NPV**	
RAD	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
ARC	88	33	59	33		91		91
pyConText	83	100	43	43	78	82	96	100
TIU	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***	***Train***	***Test***
ARC	23	0	100	0		100		84
pyConText	*78*	100	58	67	92	90	96	100

Conclusions: ARC can filter negative RAD carotid images with high NPV, but not for TIU which has a higher prevalence of negative carotid images. In contrast to ARC, pyConText filters true negative stenosis cases (100% Test NPV), and retains all positive stenosis cases (100% Test Sensitivity) for both report types. In future work, we will increase our sample size, evaluate a combined system approach for filtering non-carotid images and negative stenosis cases, and estimate chart review efficiencies on big data with this combined system approach.

Acknowledgements: VA HSR&D Stroke QUERI RRP 12-185, NIH NHLBI 1R01HL114563-01A1, & NIGMS R01GM090187, Department of Veteran Affairs, & University of Utah Institute Review Board.

References:

1. D'Avolio L, Nguyen T, Fiore L. The Automated Retrieval Console (ARC): Open Source Software for Streamlining the Process of Natural Language Processing. Proceedings of the Association of Computing Machinery's International Health Informatics Symposium: 2010 Nov. 11-12; Arlington, VA; p. 469-473.
2. Chapman B, Lee S, Kang H, Chapman WW. Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm. J Biomed Inform 44 (2011) 728–737.
3. Mowery DL, Franc D, Ashfaq S, Zamora T, Cheng E, Chapman WW, Chapman BE. Developing a Knowledge Base for Detecting Carotid Stenosis with pyConText. AMIA Symp. Proc. 2014. Washington DC. 1523.

An Approach for Partial Automation of Next Generation Sequencing Trio Analysis

Amanda J. Murphy, BS[1]; Christine Henzler, PhD[2]
[1]University of Minnesota Medical School, Minneapolis, MN;
[2]Minnesota Supercomputing Institute, University of Minnesota Minneapolis, MN

Abstract
We have written a script for the University of Minnesota Medical Center clinical inherited disease analysis pipeline that will work with separate annotated variant call files to partially automate analysis and reduce the potential for human error.

Introduction

Massively parallel sequencing, or next generation sequencing (NGS) is a powerful tool available for physicians and researchers diagnosing and studying inherited disease and cancer. However, a number of challenges still exist in optimizing NGS for routine use in clinical practice. In particular, filtering, prioritizing and interpreting variants identified through NGS pipelines still frequently involves manual analysis. Automating these steps increases efficiency and reduces the potential for human error.

Methods

Using the R programming language, we have created a script that partially automates the analysis of inherited disease sequences across trios, defined as an affected patient and his/her two biologic parents, as well as non-standard trios (a trio where one or both parents are replaced by other biologic relatives) and larger sets of relatives. To the best of our knowledge, this is the first non-commercial script that can take an unspecified (within memory allocation restraints) number of variant call files for related individuals, merge the data and produce a single output file that categorizes potential disease-causing variants based on inheritance patterns. The mode of inheritance (autosomal dominant, autosomal recessive, or X-linked recessive) is often not known, so variants are filtered and prioritized under all three modes of inheritance, and *de novo* mutations are also identified. Since variants with allele frequencies in the general population >1% are unlikely to cause a rare inherited disease, we filter as lower priority, but do not eliminate, data with an allele frequency >1%. To run the script, a user first creates a case-specific input file listing the vcfs for all individuals as well as their disease status (affected/unaffected), gender, and lineage (maternal/paternal) in relation to the patient. The script then uses these categories to sort and prioritize the patient's variants for review by a geneticist.

Results

A single Excel file is generated by the script. This output file contains tabs, each of which contain variants (categorized based on genotypes found in the separate files) into the following modes of inheritance: autosomal dominant, autosomal recessive (compound heterozygous), and X-linked recessive. If a variant is found in the patient, but does not match any variant found in the biologic relatives, it is categorized into a separate tab for potential *de novo* mutations.

Discussion

Because variants are filtered to identify only those that are consistent with the data from all individuals and each inheritance pattern, as well as prioritized by allele frequency, analyzing NGS clinical data is made more efficient. We have created this script while in consult with pathologists and clinical geneticists, and hope to incorporate this script into our inherited disease NGS pipeline within the next year.

Pathways analysis of rare, high-impact variants in mothers of children with autism

Chloe O'Connell[1], Sasha Sharma[2,3], Jae-Yoon Jung[2,3], Dennis P. Wall[2,3]
[1] Stanford Medical School, [2] Division of Systems Medicine, Stanford University, [3] Hartwell Autism Informatics Initiative (iHART)

Studies have linked autism to numerous environmental factors during pregnancy, including exposure to pollution, viral infection, and psychological stress. Here we investigate the possibility that rare variants may interact with environment to influence the birth risk of autism. Using DAVID, we performed a pathways analysis of rare, high-impact variants in mothers of children with autism. We found that pathways involved in oxidative stress and immune response were significantly enriched for variant-containing genes in the mothers and not the fathers of autistic children.

Background

Events during pregnancy, such as maternal infection[i], stressful life events[ii], and even exposure to severe storms[iii] and maternal residential proximity to areas of agricultural pesticide application[iv], have been linked to increased risk of giving birth to a child with autism. However, the effect of maternal genotype and its interaction with these events is unknown. The few existing studies on the topic have searched for single alleles that impact autism risk in the fetus, rather than the pathways to which these alleles belong. A putative role for the GSTP1*G313A haplotype of the glutathione S-transferase gene GSTP1 has also been suggested[v]. GSTP1 is an enzyme used to conjugate molecules to glutathione, detoxifying cell environments. In addition, a RFC1 allele, encoding a reduced folate carrier protein responsible for the delivery of 5-methyltetrahydrofolate into cells, has also been implicated as a potential "teratogenic allele"[vi]. However, a more recent GWAS study failed to discover any variants of genome-wide significance[vii].

Methods

Filtering variants to only those found in <1% of the samples in the 1000 genomes project, we developed a list of rare alleles in 239 unique mothers of children with autism, within 206 exomes and 33 whole genomes. We further filtered the rare variants via SnpEff predicted impact, selecting only those with a predicted impact of "high". We then compiled a list of 591 genes containing rare, high-impact variants in mothers of children with autism, and tested for pathway enrichment.

Results

We discovered 4 pathways enriched for variants in the mothers' genomes in the Kegg and Reactome pathway databases. Of note, the Kegg pathway "Metabolism of xenobiotics by cytochrome P450" was significantly enriched for variants in the mothers of children with autism (p = .0216*), but not the fathers (p = 1.00*) or the probands themselves (p = 1.00)*. The most significantly enriched Reactome pathway in the mothers, "Biological Oxidations" (p = 3.36E-04*), also was not enriched in the fathers of trios (p = 1.00*). It was, however, significantly enriched in the probands themselves, although less so than in the mothers (p = .041*). In fact, none of the pathways in the Kegg or Reactome databases remained significantly enriched in the list of variant-containing genes in fathers of children with autism after Bonferroni correction.

* = indicates p-value is Bonferroni corrected

Discussion

Given their roles in protection from oxidative stress, particularly xenobiotic-induced stress, the genes within the pathways found to be enriched within autism mothers are worth further investigation. Furthermore, our results are consistent with the hypothesis that toxic and oxidative stress may impact autism risk during fetal development. More investigation is required to determine the nature of these potentially teratogenic variants, as well as the nature of the gene-environment interactions they may influence.

[i] Atladóttir et al. *J. Autism Dev Disord.* **2010**, 40, 1423.
[ii] Beversdorf et al. *J Autism Dev Disord.* **2005**, 35, 471–478.
[iii] Kinney et al. *J Autism Dev Disord.* **2008,** 38, 481.
[iv] Shelton et al. *Environ Health Perspect.* **2014**, 1–4.
[v] Williams et al. *Arch Pediatr Adolesc Med.* **2007,** 161, 356.
[vi] James et al. Am **J** Med Genet B Neuropsychiatr Genet. **2010,** *153*, 1209.
[vii] Yuan & Dougherty. *Autism Res.* **2014**, 7(2), 245.

Cancer Registry Control Panel (CRCP): A System for the Discovery of Patient Cancer Status using the EMR

John D. Osborne MS, Matt C. Wyatt MS, Andrew O. Westfall MS, James H. Willig MD, Geoff Gordon MSEE

University of Alabama at Birmingham, Birmingham, Alabama

Summary / Abstract

CRCP assists cancer registrars in identifying reportable cancer cases for abstraction. By using ICD9 codes and NLP-identified positive cancer mentions from pathology reports we can identify candidate cancer patients with higher precision and speed than the unassisted manual process while increasing abstracted cancer patient volume by 41%.

Introduction and Background

As one of 45 states participating in the National Program of Cancer Registries (NPCR) Alabama mandates that UAB and other cancer care providers report cancer cases to the state. Historically registrars at UAB identified cancer patients by manual inspection of printed pathology reports. We replaced this process with CRCP, a system that automatically retrieves pathology reports from the EMR and detects cancer mentions using Natural Language Processing (NLP) techniques. Cancer related ICD9 codes are also used to increase both the precision and recall of the system.

Methods

Cancer mentions are detected using a UIMA-AS[1] pipeline containing NLM's UIMA MetaMap[2] annotator and a variety of custom rule-based UIMA annotators that primarily act to filter MetaMap hits. Filtering is done on the basis of document segmentation (regular expression based), semantic content (by CUI and UMLS semantic type), presence of negation, patient subject class (excluding cancer mentions when they refer to family members) and cancer reportability status based on complex rules (updated yearly by the state) that specify whether a cancer is reportable or not (for example most benign tumors are excluded). CRCP inspects pathology reports nightly to identify pathology records containing relevant cancer concepts and combines this with ICD9 codes from the EDW to identify candidate cancer patients. Cancer concepts are then highlighted in all candidate clinical notes and sorted in CRCP for faster registrar candidate validation and abstraction.

Results

The results of reportable cancer patient detection by NLP on pathology documents alone, ICD9 code alone and combined NLP and ICD9 data for a 3 month window are displayed in Table 1. Table 2 shows the results of a more recent CRCP cancer patient detection algorithm with higher precision Also indicated is a 41% higher overall year over year abstraction rate for a comparable 2 month reporting period.

The increase in throughput is due in part to the fact that 51.2% of incoming patients (from pathology reports) are triaged as unreviewable since the NLP process has failed to identify any cancer concepts - significantly reducing registrar review burden. Additionally, the unstructured nature of the old manual process resulted in 27% of patients being reviewed by more than one registrar and 10% of PTH documents being reviewed more than once. CRCP eliminated duplicate patient and pathology document review completely.

Finally, the CRCP interface displays cancer concepts immediately after registrars click on a note, thereby reducing the note review burden which can reduce abstraction time. Despite this reduction in review, the total number of discovered cancer patients is now greater in the CRCP system due to the inclusion of ICD9 codes in the detection process.

Conclusion

CRCP details the benefits of an automated cancer patient detection over a manual process and quantifies those benefits in terms of abstracted patient throughput, recall and precision. We also demonstrate that a combination of ICD9 codes and NLP gives the highest performance in cancer patient discovery. Under development is an improved cancer patient detection algorithm based on machine learning using CTAKES[3] 3.2 and the ClearTK[4] toolkit.

Tables

Table 1. Precision of Reportable Cancer Detection for a 3 Month Window

	NLP Only Cases	ICD9 Only Cases	Combined
Accepted	955	1061	912
Rejected	1429	1117	333

Table 2. Performance of Previous PTH Based System versus CRCP

System	Case Finding Accuracy	Average Monthly Throughput
Paper System (PTH)	16.2	380.5
CRCP (PTH+ICD9)	77.8	537.5

References

1. Getting Started: UIMA Asynchronous Scaleout. The Apache Software Foundation. [cited 2014 September 25]; Available from: http://incubator.apache.org/uima/doc-uimaas-what.html.
2. Aronson, AR. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In *Proc AMIA Sym.* p. 17. American Medical Informatics Association, 2001.
3. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC,Chute CG. "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications." *JAMIA* 17, no. 5 (2010): 507-513.
4. Ogren PV, Wetzler PG, and Bethard SJ. "ClearTK: A UIMA toolkit for statistical natural language processing." *Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP* 32 (2008).

Automating i2b2 Ontology Mappings with MetaMap

Lori C. Phillips, MS[1], Shawn N. Murphy, MD, PhD[1,2]
[1]Partners HealthCare, Charlestown, MA, [2]Massachusetts General Hospital, Boston, MA

Abstract

In patient query systems such as i2b2, the norm is to query against a known standard vocabulary such as ICD9 for diagnoses, LOINC for lab tests, etc. As a result there is often a need to map local EHR codes to a standard vocabulary. We describe here an extension to i2b2's mapping tools whereby the UMLS MetaMap is accessed dynamically to generate mappings.

Background

In previous work, we developed a set of Ontology Mapping tools within the i2b2 framework. The tools did not attempt to automate the mapping process, but rather, provided a mechanism for users to either map terms manually or verify third party mappings. We extend these tools by linking them to MetaMap[1], a tool which maps biomedical terms to selected UMLS vocabularies.

Methods

Central to the original mapping tools is a collection of terms to be mapped or merged into a destination hierarchy. These terms are typically shown in tabular format and may be assigned to a location in the destination tree via a drag and drop user interface. We extended the mapping tool by linking it to MetaMap via a series of API calls. Each unmapped term is sent to MetaMap which returns a suggested mapping if one exists. Users are then allowed to accept or reject the mappings returned.

Results

We tested the platform using 1598 terms that originate from several sources within Partners and all derive from non-standard coding systems. Our goal was to use MetaMap to map them to our local i2b2 2014 version AA of ICD9CM. Of the 1598 terms, 702 were mapped to ICD9 by MetaMap. Of the 702 mapped terms, 10 were incorrect; 138 were generalized ("Fracture" v "Radius fracture"); the remaining 554 were correct. Additionally, MetaMap identified another 187 terms but returned ICD9 descriptions not found in our local ICD9 instance ("Tachycardia NOS" vs "Tachycardia, unspecified").

Conclusion

We have built an extension to the i2b2 Ontology Mapping tools to automate the mapping and placement of local terms within destination UMLS ontologies. The tool is interactive in nature allowing users to accept or reject the mappings returned by MetaMap.

This work was funded by U54HG007963.

References

[1] AR Aronson and FM Lang, An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229,2010.

Definition and Exploration of LUAD-LUSC Hybrid Subtype via Integrative Omics Module Networks

Katie Planey, MBS, MS[1], Olivier Gevaert, PhD[1]
Stanford University

Introduction

While lung adenocarcinoma (LUAD) and lung squamous (LUSC) carcinoma are considered clinically distinct lung cancers, we discovered a TCGA pancancer methylation cluster spanning a significant portion of LUAD and LUSC samples. We hypothesized that this hybrid cluster is driven by common cancer driver genes and used multi-omics data fusion to identify them.

Methods

Our module network methods integrate three omics datasets: RNA-seq gene expression, methylation, and copy number variation. We used RNA sequencing, DNA methylation and copy number data of 571 LUAD and 535 LUSC cases. RNA-seq paired-end fastq files were aligned using the STAR aligner with the hg19 reference genome[..]. Methylation data was prepared by first clustering probes on the Illumina 450k platform by genes, and then these genetic feature clusters were fed into MethylMix, a method that models methylation states using density mixture models and determines which genes are significantly hyper or hypomethylated in cancer vs. normal tissue samples (Gevaert et al, in press.) To discover genes that have significantly amplified or deleted CNV, we used the GISTIC 2.0. Significant genes identified from MehylMix and/or GISTIC were considered potential regulatory genes..We then used AMARETTO to derive cancer driver genes and connect them with their targets.

AMARETTO is an algorithm to identify cancer drivers by integrating a variety of omics data from cancer data. AMARETTO accomplishes this using a multi-step algorithm that integrates copy number, DNA methylation and gene expression data to identify cancer driver genes and subsequently associates them with their downstream targets through module network analysis.
To make the resulting gene regulatory modules from AMARETTO directly therapeutically, we only used cancer driver genes from the druggable genome AMARETTO was run with the newly identified methylation-based LUSC-LUAD hybrid cluster of 418 patients; including 155 LUAD and 263 LUSC cases.

Results

The average R squared across all modules was .6 and 5 regulatory-nonregulatory modules had an R squared above .75. The module with the highest R squared and R squared adjusted of .79 contained 10 druggable regulatory genes and 97 nonregulatory genes; we plan to further analyze this gene module to better understand how one might develop novel clinical regimens for this LUAD-LUSC subtype.

Discussion

We thus hypothesize that this group of mixed LUSC and LUAD patients does constitute biologically distinct signals, and that the majority of LUSC and LUAD patients cannot be assumed to be biologically homogeneous.

Biocuration of chemopredictive markers and direct drug targets from public data to determine associated targeted therapy benefit and clinical outcomes

Shruti Rao, MS[1], Shahla Riazi, MD, PhD[1], Simina Boca, PhD[1], Varun Singh, MS[1], Michael Harris, MA[1], Peter McGarvey, PhD[1], Michael J. Pishvaian, MD, PhD[2], Jonathan Brody, PhD[3], Subha Madhavan,PhD[1]
[1] Innovation Center for Biomedical Informatics, Georgetown University
[2] Department of Hematology/Oncology, Lombardi Comprehensive Cancer Center, Georgetown University
[3] Department of Surgery, Thomas Jefferson University

Summary
Advances in cancer biomarker research have improved early disease detection and provided guidance on choosing appropriate individualized therapies. Our goal was to extract, standardize, and organize molecular information and treatment options from personalized medicine related publications that can ultimately be used to understand clinical utility of these biomarkers.

Introduction and Background
The selection of personalized cancer therapy based upon a patient's molecular profile requires an enormous amount of data wrangling to collect, review, analyze and integrate molecular, clinical, patient-specific history and pharmacological data. Tumor boards are constantly faced with challenges in making therapy decisions based on empirical evidence. In an attempt to address this issue, our goal was to extract, standardize and organize molecular information and treatment options from personalized medicine related publications. We first focused on the protein expression status of chemo-predictive biomarkers that also currently lack sufficient evidence for clinical use. Once curated, the structured data can be used to generate novel scientific hypotheses, design new studies, obtain a better understanding of biological mechanisms of disease, perform meta-analyses, and create clinical decision support systems. Our efforts support the paradigm shift from focusing on choosing drugs based on diseases to choosing drugs based on biomarker status for a particular disease or, in some cases, based solely on molecular biomarkers.

Methods
We used PubTator, a web based text mining tool, to extract relevant PubMed articles about chemo-predictive biomarkers and associated therapies. We manually curated 87 PubMed articles on ERCC1 and platinum-based therapies. Each article was organized into three broad categories: 1) disease type and therapy, 2) biomarker information, and 3) study information. Data was manually curated and standardized within each of these categories to determine the predictive effect of biomarkers on therapeutic outcomes. We then identified study types and assigned evidence levels ranging from I-V to these articles. Well-designed randomized controlled trials were assigned a high level (I) of evidence whereas in-vitro studies, expert opinions and single patient case studies were assigned lower levels (IV-V) of evidence.

Results and Discussion
Several studies have suggested that the excision repair cross-complementation group 1 (ERRC1) gene can predict the clinical benefit from platinum-based chemotherapy. Our results show that, in general, low expression levels of ERCC1 are associated with benefit of platinum-based therapies. However, evidence levels supporting this association vary for different cancer types. For lung cancer, more level I and II studies show benefit of low ERCC1 expression on platinum-based therapies whereas for pancreatic cancer, there is only level III evidence demonstrating benefit. Curation of chemo-predictive biomarkers TUBB3, TS, TOPO1, RRM1, EGFR, BRCA2 and cMET are ongoing and will ultimately be made available through a public resource.

Conclusion
We discovered that peer reviewed studies that involved predictive biomarkers had: 1) a diversity of methods used; 2) at times, an inconsistency in conclusions; and 3) clinical information based on the tumor type (e.g., lung vs pancreas). The organization and analysis of dispersed public data for retrospective biomarker analysis and other associated metadata enables researchers to readily generate hypotheses for new clinical trials and to explore the use of published markers to stratify patients upfront for 'best-fit' therapies.

Definition and Application of Phenotype Design Patterns

Luke V. Rasmussen[1], Will K. Thompson[2], Jennifer A. Pacheco[1], Abel N. Kho[1], David S. Carrell[3], Jyotishman Pathak[4], Peggy L. Peissig[5], Gerard Tromp[6], Joshua C. Denny[7], Justin B. Starren[1]

[1]Northwestern University Feinberg School of Medicine, Chicago, IL; [2]NorthShore University HealthSystem, Evanston, IL; [3]Group Health Research Institute, Seattle, WA; [4]Mayo Clinic, Rochester, MN; [5]Marshfield Clinic Research Foundation, Marshfield, WI; [6]Geisinger Health System, Danville, PA; [7]Vanderbilt University, Nashville, TN

Abstract

Design patterns, in the context of phenotype algorithm development, provide generalized guidance on the interpretation, use and combination of strategies to account for nuances in electronic health record (EHR) data. They provide generalized descriptions of approaches to take, and as such are agnostic to any specific content, EHR or technology platform. Based on a review of algorithms developed by the electronic Medical Records and Genomics (eMERGE) Network, we propose 21 phenotype design patterns, and provide guidance for their application to future algorithms.

Introduction

The increased implementation of electronic health records (EHRs) has offered new avenues for institutions to exploring more in-depth EHR-based phenotype algorithm development[1]. Given the multiple modalities in which data are collected in the EHR, differences in how data are collected across institutions, and the nuances and surrounding context of that data collection, phenotype algorithm development is not a trivial task[2]. A library of phenotype design patterns may help future algorithm authors understand these nuances sooner. We report on the methods and results of an in-press publication to describe such a library[3].

Methods

A review of 24 EHR-based phenotype algorithms developed by the eMERGE network was conducted to assess common themes in the algorithm artifacts. The recurring themes, coupled with our collective experiences in developing phenotype algorithms, were used to construct a list of phenotype design patterns.

Results

In total, 21 phenotype design patterns of varying levels of complexity were identified. The pattern definition includes a description and justification for the pattern, recommendations on when to use and possibly when not to use it, as well as examples of the pattern in practice. One of the more common patterns, "Rule of N", states that requiring a minimum number of occurrences of an event may increase the probability that a patient does or does not have a condition. This approach can be used to correct for the use of billing diagnoses, where a diagnosis code appears for screening or ruling out of a condition. A consideration with this pattern is that a large N may indicate the algorithm is over-fitting to the local data set.

Discussion

The identification and curation of EHR-based phenotype design patterns provides a new source of knowledge to inform both new and experienced phenotype algorithm developers. While no formal taxonomy of these patterns is presented, we have observed independence and interdependence of many patterns. This assessment provides additional insight for future phenotype authors to aid in the appropriate selection and use of patterns in future work.

Acknowledgement. This work funded by NHGRI grants: U01HG006389, U01HG006382, U01HG006375, U01HG006379, U01HG006388 and U01HG006378.

References

1. Hripcsak G, Albers DJ. Next-Generation Phenotyping of Electronic Health Records. Journal of the American Medical Informatics Association. 2013;20(1):117-21.
2. Denny JC. Chapter 13: Mining electronic health records in the genomics era. PLoS Comput Biol. 2012;8(12):e1002823.
3. Rasmussen LV, Thompson WK, Pacheco JA, Kho AN, Carrell DS, Pathak J, et al. Design patterns for the development of electronic health record-driven phenotype extraction algorithms. J Biomed Inform. 2014;(In press).

RED-i Data Integrator for REDCap Projects

Erik C. Schmidt, BS, MCSE, A+, CSM, Christopher P. Barnes, William Hogan, MD, PhD, Josh Hanna, MS
University of Florida, Gainesville, FL

Abstract

RED-I (Research Electronic Data Integrator) is a data-driven software package that uses XML trees to translate and transform data elements from a data source (such as an EMR system) to a REDCap project in an automated fashion. RED-i utilizes component mapping, making use of standard terminologies such as LOINC for lab data, SNOMED for clinical terms, and RxNorm for medications. The RED-i project was developed at the University of Florida, has been tested to work with various EMR systems at multiple institutions, and has been released as an open-source project under a BSD 3-Clause license.

Introduction

In this age of information, manual processes have become a very inefficient way to meet the data needs of research projects. The convergence of open source software and subject matter expertise now make improved information management an attainable goal for patient studies. The Clinical and Translational Science Informatics and Technology (CTS-IT) group at the University of Florida has developed a system that automates the request, delivery, and import of EMR data directly into a REDCap project, saving countless hours of manual data entry and preventing data transcription errors.

Methods

Starting with a REDCap project and making use of the built-in API, we were able to develop a suite of tools, using Python, which performs the following operations: (1) pulls elements for an EMR data request from a REDCap form, (2) translates the form data into a prescribed format, (3) delivers the formatted request to an EMR system, (4) returns the requested EMR data in a prescribed format, (5) translates the EMR data into a standard such as LOINC, and (6) finally writes the EMR data directly to a form within the REDCap project.

Results

We have accomplished all original goals of the project and are successfully delivering EMR data to a REDCap project in a completely automated fashion. This data request and delivery process is taking place on a scheduled, daily basis from two different Epic EHR systems, and one Cerner EHR system, and can deliver data to any REDCap system for which we have an API key available to us.

Discussion

Interviews with our customers indicate that a process which originally took approximately 10 hours of manual data entry is being accomplished in less than 10 minutes of manual entry with our software in place. In addition, logic would dictate that errors in research data due to the manual entry process will be reduced, limiting errors to those that might exist in the EMR data itself.

Learning Objectives

After participating in this session, the learner should be better able to:

- Reduce the time spent with manual data collection and entry into a research system, such as REDCap
- Reduce the occurrence of data transcription errors in research data
- Use research data to collaborate with other researchers or institutions in a standards-based fashion
- Minimize the occurrence of unnecessary viewing of data, in particular PHI

A Comprehensive Time-Course-Based Meta-Analysis of Sepsis and Sterile Inflammation Reveals a Robust Diagnostic Gene Set

Timothy E. Sweeney, MD, PhD[1,3], Aaditya Shidham[3], Hector R. Wong, MD[4,5], Purvesh Khatri, PhD[2,3]

[1] Department of Surgery, Stanford University School of Medicine

[2] Stanford Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine

[3] Stanford Center for Biomedical Informatics Research, Stanford University

[4] Division of Critical Care Medicine, Cincinnati Children's Hospital Medical Center and Cincinnati Children's Research Foundation

[5] Department of Pediatrics, University of Cincinnati College of Medicine

Abstract

Six publically available gene expression cohorts comparing SIRS/trauma to sepsis were split into time-matched sub-cohorts and summarized via meta-analysis, which yielded an 11-gene set that was validated for discriminating patients with sepsis or infection in 16 independent cohorts.

Introduction

Although several dozen studies of gene expression in sepsis have been published, distinguishing sepsis from a sterile systemic inflammatory response syndrome (SIRS) is still largely up to clinical suspicion. We hypothesized that a time-course-based multi-cohort analysis of the publically available sepsis gene expression datasets would yield a robust set of genes for distinguishing patients with sepsis from patients with sterile inflammation.

Methods and Results

A comprehensive search for gene expression datasets in sepsis identified 27 datasets matching our inclusion criteria. We chose five datasets (n=663 samples) comparing patients with sterile inflammation (SIRS/ICU/trauma) to time-matched patients with infections. We applied our multi-cohort analysis framework that utilizes both effect sizes and p-values in a leave-one-dataset-out fashion to these datasets. We identified 11 genes that were differentially expressed (FDR cutoff ≤ 1%, inter-dataset heterogeneity p>0.01, summary effect size >1.5 fold) across all discovery cohorts with excellent diagnostic power (mean AUC 0.87, range 0.7-0.98). We then validated these 11 genes in 15 independent cohorts comparing (1) time-matched infected vs. non-infected trauma patients (4 cohorts, 218 samples; mean AUC 0.83, range 0.73-0.89) (2) ICU/trauma patients with infections over clinical time-course (3 datasets, 215 samples, AUC range 0.68-0.84), and (3) healthy subjects vs. sepsis patients (8 datasets, 446 samples; mean AUC 0.98; range 0.94-1.0). In the discovery Glue Grant cohort, SIRS plus the 11-gene set improved prediction of infection (compared to SIRS alone) with a continuous net reclassification index of 0.90. Using deconvolution methods and cell-type specific signatures, we show that the gene set may be representative of cell-type shifts in infection.

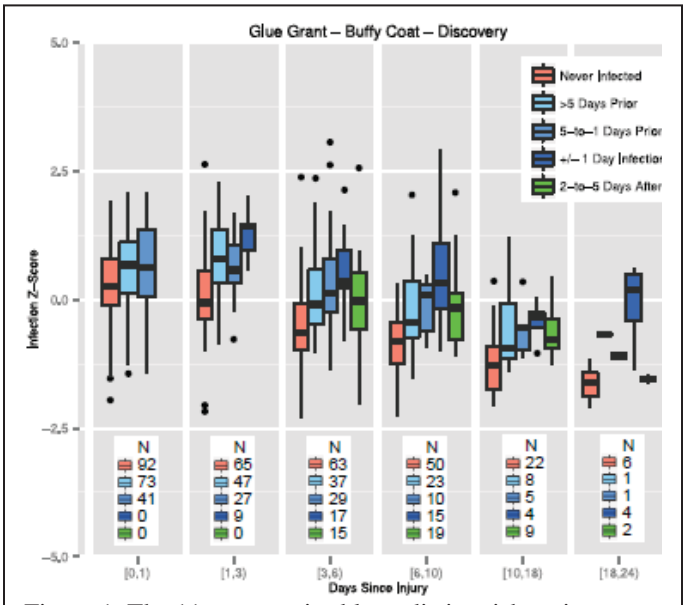


Figure 1. The 11-gene set is able to distinguish patients at the time of diagnosis of infection from matched non-infected patients. The 'infection score' begins to rise prior to clinical diagnosis.

Conclusion

Overall, multi-cohort analysis of time-matched cohorts yielded 11 genes that robustly distinguish sterile inflammation from infectious inflammation.

Improving Phenotypic Granularity with NLP-PheWAS in Multiple Sclerosis

PL Teixeira, MS[1]; MF Davis, MS, PhD[2]; LK Wiley, MS[1]; L Bastarache[1]; JC Smith, MS[1]; RJ Carroll, PhD[1]; D Fabbri, PhD[1]; DM Roden, MD[1]; JC Denny, MD, MS[2]

[1]Vanderbilt University, Nashville, TN; [2]Brigham Young University, Provo, UT

Introduction and Background: Phenome-wide association studies (PheWAS) identify associations between a genotype and many phenotypes, typically using billing codes. We have recently shown that PheWAS on a population of 6,260 individuals using NLP-derived phenotypes (NLP-PheWAS) from patient problem lists (PL), summaries of patient visits (discharge summaries DS), and clinic notes that document the history and physical (H&P), can replicate and sometimes improve upon ICD9-based method. Here we show the power of NLP-PheWAS to identify more granular phenotypes for 25 known multiple sclerosis (MS) risk variants in a 4.7 fold larger population.

Methods: We processed 1,877,111 inpatient and outpatient clinical notes from 29,685 European-American individuals (ancestry determined using STRUCTURE) from BioVU, Vanderbilt's biobank linked to EHR. As described previously, we mapped content from high value sections to SNOMED-CT concepts with negation status, note section, and experiencer information using the KnowledgeMap Concept Indexer (KMCI) and SecTag. We removed concepts that were negated, possible, or for individuals other than the patient and limited to the following UMLS semantic types: findings, diseases or syndromes, therapeutic or preventive procedures, signs or symptoms, neoplastic processes, pathologic functions, and congenital abnormalities. We studied 8,943 phenotypes occurring in at least 25 individuals. Using the NHGRI GWAS catalog, we selected 78 SNPs related to MS with p-value < $5*10^{-8}$, 25 of which were available on our genotyping platform. We performed logistic regression of additively encoded genotypes adjusting for age, gender, and original dataset using PLINK v1.90a 64-bit. We reviewed associated phenotypes with p-values < 0.01 for concepts associated with MS and grouped the resulting concepts, displaying a heat map of the odds ratio (OR) for the most significant result in each group.

Results and Discussion: One SNP, rs3135388 in HLA-DRA, had 3 associations meeting the Bonferroni correction of p<2.24e-7: relapsing-remitting MS (p=3.1e-19, gene=HLA), secondary progressive MS (1.7e-13, HLA), and muscle spasticity (2.8e-9, HLA). Additional interesting associations in the HLA region that did not meet Bonferroni significance include myopathy (1.1e-4), progressive relapsing MS (2.1e-4) exacerbation of MS (8.2e-3 IL2RA) and primary progressive MS (3.6e-2). In addition, we also detected many autoimmune associations - insulin dependent diabetes mellitus (2.8e-23 HLA), dermatitides (3.6e-6 HLA), and rheumatoid arthritis (1.6e-4 HLA). Figure 1 shows the many associations found across groupings of MS-associated phenotypes (aggregated by chromosome). We grouped phenotypes based on these categories and visualize each group's most significant odds ratio (OR) in the heat map (blue - OR<1, gray - OR=1, and red - OR>1). Odds ratios range from 0.43-3.1. Four of the loci had an association with a multiple sclerosis specific phenotype, but all included several associations with known MS symptoms and Figure 1 demonstrates that different loci are associated with different MS symptomology and MS subtypes. For example, IL2RA (chr10) was associated with MS exacerbations (p=0.008) but not other forms of MS at p<0.05 The HLA region was associated with risk of all MS symptoms except visual impairment and weakness while others demonstrated a more mixed pattern. In summary, we find that NLP-PheWAS replicates known associations with greater granularity than billing codes. Such studies may aid in dissecting variable disease presentations such as seen in MS.

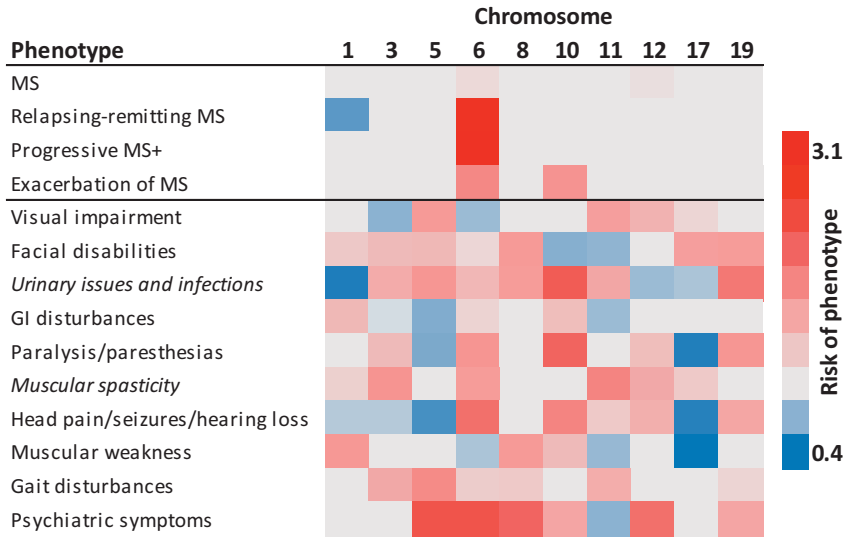
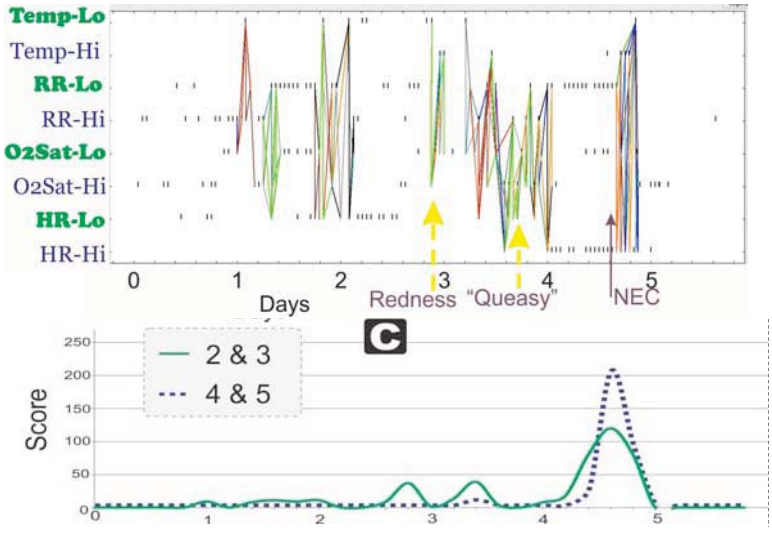


Figure 1: Multiple Sclerosis Risk Variant-Phenotype OR Heat Map

Big Data for Little Babies: Early Prediction of Adverse Events in the Neonatal Intensive Care Unit

KP Unnikrishnan, PhD, and Sidhartha Tan, MD
NorthShore University HealthSystem, Evanston, IL, USA

Summary: Few studies have mined Big Data in the Neonatal Intensive Care Unit (NICU), especially of all the vital signs. We present a case study of early prediction of necrotizing enterocolitis (NEC) through Temporal Data Mining of vital signs in the NICU.
Introduction: Death or major morbidity affects 50-60% of babies in the NICU and one in ten of low birth weight babies die. Hence early identification of clinical problems in the NICU has the most beneficial impact on patient outcome, much more so than in older children, adults, and the elderly. Initial attempts have emphasized one biological parameter, heart rate. Few studies have investigated all the vital signs of a patient (heart rate, respiratory rate, oxygen saturation, and temperature) through Data Mining. Temporal Data Mining (TDM) is powerful tool for discovering implicit, previously unknown, and potentially useful sequential information from time-stamped data. It also allows computationally efficient discovery of statistically significant sequences and event-combinations that naturally capture the data generator's time-evolution.
Methods: From data that contains ordered pairs (E_i, t_i), where E_i denotes the event type and t_i is the time of its occurrence, TDM methods discover all statistically significant sequences For example, ($E_K \rightarrow E_P \rightarrow E_U$) is a sequence where an event type E_K is followed some time later by E_P and then E_U, in that order. The algorithms stop after discovering the longest set of significant sequences.
Results: In a case study patient, the four variables were collected from electronic medical records and binarized as one sigma above or below the mean. The longest, statistically significant sequences (with 6 events) occurred more than a day before the diagnosis of NEC, at the same time when one of the experienced nurses felt "Queasy" about the baby. Another set of significant 5-event-long sequences occurred co-incident with the NEC diagnosis. 2, 3, and 4-event long sub-sequences of these occur up to 3 days prior to the NEC diagnosis and provide feature sets for early prediction.



Top figure: All significant 5-event patterns (color coded) plotted on binarized data. Note that pattern sets occur near significant physiological events marked on x-axis.

Bottom figure: A score from 5-event patterns that occur during NEC diagnosis and their 4-event long sub-patterns show high specificity and sensitivity at time of diagnosis (dashed blue line). The score from 2 and 3 event long sub-patterns show peaks earlier and can be used for early prediction of NEC.

Discussion: Temporal Data Mining is a useful method for early prediction of adverse events in the NICU.

Personalized Medicine on Cancer Treatment: Using Big Data Techniques to Integrate Clinical and Genomic Data

Filippo Utro, PhD, Ping Zhang, PhD, Fei Wang, PhD, Jianying Hu, PhD
IBM T.J. Watson Research Center, New York, USA

Summary

In the last few years genomic data are becoming largely available and their use in the context of personalized medicine is going to be a standard de facto. We propose a new method to identify possible drug recommendation using genomic data as well as clinical information of the patient. We show that our approach is outperforming the competitors in a real case scenario using public cancer data.

Introduction

In contrast to the one-size-fits-all medicine, personalized medicine aims to tailor treatment to the individual characteristics of each patient. Personalized medicine will allow targeted prescription due to the specific profile of the patient avoiding adverse reactions and expensive treatments that will not be effective. In this study, we propose a joint matrix factorization (JMF) approach which can be used to generate personalized drug recommendations by integrating multiple drug and patient similarity networks.

Methodology

The JFM method is applied to generate personalized drug recommendations. JMF formulates the task of personalized drug recommendation as a constrained non-convex optimization problem. It utilizes multiple drug similarity networks, multiple patient similarity networks, and known effective drug-patient associations to explore potential new associations among drugs and patients with no known links. The method was proposed for a drug repositioning task[1], here we adopted it in the personalized medicine scenario. We used the Tanimoto coefficient (TC) to measure the drug and patient similarities. In particular the chemical structure information is used for the pairwise drug similarity, while the patient similarities is computed using the demographic (i.e. age, race and gender) and SNPs information.

Results

We used Glioblastoma multiforme data from TCGA, and we define treatments for patients who live for more than 15 months (i.e. median survival with standard-of-care) as effective. We extracted patients that have clinical information (i.e. treatment information), genomic information (i.e. SNPs), and have a positive response to treatment, resulting in a cohort of 118 patients. The 118 patients took one or more drugs. We used a leave-one-out cross-validation scheme to evaluate treatment recommendation algorithms. We considered four treatment recommendation methods: (1) Label propagation (LP) using only patient demographic information. (2) LP using only patient SNPs information. (3) LP using only drug information. (4) Our proposed JMF method by considering drug information, patient demographic and SNPs information.

Table 1 shows that (i) our method is effective in integrating multiple drug/patient similarities for treatment recommendation tasks; (ii) the combination of clinical (e.g. demographics), genomic (e.g. SNPs) and drug (e.g. chemical structures) information can help to identify which drug is likely to be effective for the patient.

Table 1. Area Under the ROC Curve (AUC) of the leave-one-out cross-validation comparison of four treatment recommendation strategies

Method	AUC
LP - Patient Demographic	0.7709
LP - Patient SNPs	0.8335
LP - Drug	0.5848
JMF - All information	0.8505

References

1. Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annu Symp Proc 2014.

Data-integration and information propagation for drug discovery in triple negative breast cancer: a network-based approach

F. Vitali MS[1], L.D. Cohen MS[1], F. Mulas PhD[1], A. Zambelli MD[2], R. Bellazzi PhD[1]
[1]Dipartimento di Ingegneria Industriale e dell'Informazione, Università di Pavia, Italy;
[2]Laboratory of Experimental Oncology and Pharmacogenomics, IRCCS-Fondazione S. Maugeri, Pavia, Italy

Abstract

The integration of data and knowledge from heterogeneous sources is a key element for scientific investigations in drug discovery. We provide a computational method that automatically identifies multi-target drugs by knowledge and data integration through a network-based model. The application to breast cancer highlights the synergistic effect of drugs that may be proposed for in vitro studies.

Introduction

Multi-target drugs have been proven to achieve superior therapeutic efficacy and safety to complex diseases than the conventional mono-target drugs that are mostly marketed today. Multi-target drugs are able to comprehensively target the pathological network of a disease and to amplify the final therapeutic success due to their treatment effects by synergy. Triple Negative Breast Cancer (TNBC) is a heterogeneous and aggressive subclass whose biology is poorly understood[1]. This tumor subtype is lacking in known targets and the only treatment option is chemotherapy; thus, it is a suitable candidate for network-centric modeling and multicomponent therapeutics, providing new therapeutic views and recommendations for drug repositioning. The approach we developed integrates different data sources and may thus better define the global picture of TNBC. The integration of knowledge allows making full use of known targets, drugs and disease pathways thus leading us to develop a network-based approach that results in the identification of optimized drug combinations and core disease causative pathways.

Methods

As a first step, we implemented ourgeneral methodology to rank drugs target combinations[2] to the case of TNBC (Figure 1.A). This approach focuses on disease-specific protein networks to rank new target candidates thanks to the definition of a Topological Score of Drug Synergy (TSDS). The protein interaction network for TNBC was constructed on the basis of a recent mutational study[1], starting from genetic changes related to the disease and integrating them with the information available both in the database STRING and in the TNBC literature. The proteins and their interactions were represented by the nodes and the edges of the network, respectively. Afterwards, significant target combinations (w.r.t the TSDS score) were further augmented through the application of a data fusion approach with penalized matrix tri-factorization[3] (Figure 1.B.1). This procedure enabled to detect new interaction pairs between drugs and the selected targets by considering various data sources, i.e. Drug Bank, STRING, Disease Ontology, Therapeutic Target Database and GeneRIF. As a following step, we extracted from KEGG the biological pathways that are related to the progression of the disease and involve the drugs identified so far (Figure 1.B.2). This information was exploited to orient the edges of the network, resorting to an adapted version of a recentlydeveloped edge orientation algorithm[4], based on random walks. The edges were oriented according to the information flow (i.e. pharmacological action of specific drugs) and by considering the knowledge derived from the biological pathways. This approach allowed us to select the best drug candidates for future in vitro studies based on the analysis of the top information network modules[5] (Figure 1.C).

Results and discussion

The analysis of results demonstrates that the method could elucidate the interactions of the complex disease under study and may suggest potential drug interventions. Orienting edges of TNBC protein interaction network presented here and the analysis of information flow in the directed network has provided valuable insight regarding potential drug combinations and possible off-target effects. A particularly interesting candidate that emerged using our method was Imatinib, a well-known drug mainly used in leukemia, which may be retargeted to TNBC in combination with other drugs. In vitro tests to confirm the hypothesis are ongoing. It is worthwhile noticing that this work integrates protein interactions, transcriptional analysis, pathways to provide in a comprehensive network for newdrug discovery that can in principle be applied to any complex disease.

References

1. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao, Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012; 486(7403): 395-399.
2. Vitali F, Mulas F, Marini P, Bellazzi R. Network-based target ranking for polypharmacological therapies. JBiomedInform. 2013; 46: 876-881.
3. Žitnik, M, Janjić, V, Larminie C, Zupan B,Pržulj N. Discovering disease-disease associations by fusing systems-level molecular data. Scientific reports. 2013; 3.
4. A. Gitter, J. Klein-Seetharaman, A. Gupta and Z. Bar-Joseph. Discovering pathways by orienting edges in protein interaction networks. Nucleic Acids Res. 2011; 39: e22.
5. Stojmirović A, Yu YK. Information flow in interaction networks II: channels, path lengths, and potentials. JComputBiol. 2012; 19: 379-403.

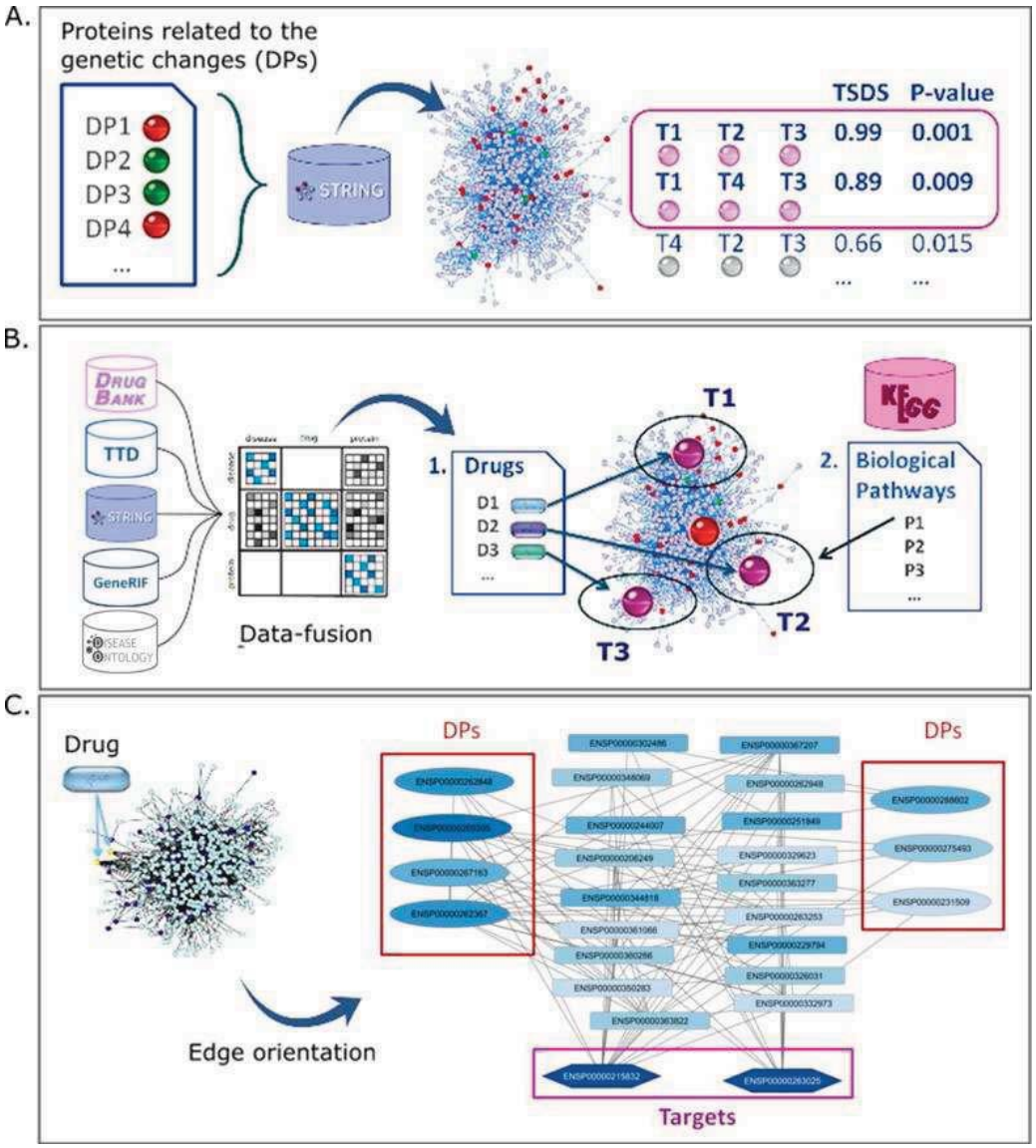


Figure 1. A) TSDS application to extract significant target combination; B) Pharmacological and biological analysis of the significant targets: 1. Identification of drug-target associations; 2. Selection of the pathways involved in the progression of the disease and related to the drug target; C) Orientation of the network edges; Analysis of information flow in the directed protein interaction network reveals a sub-module through which most of the information flows (sources of information flow (i.e. targets) are depicted as hexagons, sinks (i.e. DPs) as ovals, and others as rectangles. Blue shades indicate relative amounts of information flow in the network submodule).

Translation of Pathway-Based trans-eQTL to Human Phenotypes

LK Wiley[1], RJ Carroll[1], TL Edwards[1], J Kirby[1], L Bastarache[1], H Kuivaniemi[2], G Tromp[2], GP Jarvik[3], DR Crosslin[3], D Carrell[3], EB Larson[3], MH Brilliant[4], SJ Hebbring[4], IJ Kullo[5], J Pathak[5], JA Pacheco[6], AN Kho[6], JC Denny[1], WS Bush[7]

[1]Vanderbilt University, Nashville, TN, [2]Geisinger Health System, Danville, PA, [3]Group Health Cooperative, Seattle, WA [4]The Marshfield Clinic, Marshfield, WI, [5]Mayo Clinic, Rochester, MN, [6]Northwestern University, Chicago, IL, [7]Case Western Reserve University, Cleveland, OH

Abstract

We performed a phenome-wide association study on 9 SNPs implicated in differential expression of biological pathways with 35,609 individuals from the eMERGE Network to determine potential phenotypic consequences of pathway dysregulation. Ten SNP-phenotype associations had p-values < $5x10^{-4}$. Of these, rs425437 was associated with lip cancer ($p=5.75x10^{-6}$), for which we propose a plausible biological explanation supported by prior evidence.

Introduction

Association of single nucleotide polymorphisms (SNPs) with gene expression levels is a common approach to identify possible mechanisms underlying disease pathophysiology. Most studies examine cis-eQTL, where the SNP and gene with modified expression are located close together on a chromosome. Trans-eQTL, where the SNP and gene are distant, are less commonly tested due to the dramatic expansion of statistical tests. We have previously proposed that a single SNP may act as a trans-eQTL by inducing expression changes in known molecular pathways. This study sought to expand that hypothesis into human phenotypes by performing a phenome-wide association study (PheWAS) on SNPs that had replicating differential expression of entire molecular pathways.

Methods

Our original study investigated 2,909 known cis-eQTL SNPs by regressing the expression levels of 11,466 genes on each SNP independently. For each SNP, these expression changes were processed by Signaling Pathway Impact Analysis (SPIA), to measure both pathway perturbation and over-representation of differentially expressed genes in KEGG pathways. We tested independent sets of 210 multiethnic HapMap II samples and 466 multiethnic 1000 Genomes samples. Fifteen SNP-pathway associations representing 13 individual SNPs replicated. Of these, 9 SNPs were available in the post-quality control imputed genome-wide genetic data for 35,609 adults across 7 study sites of the Electronic Medical Records and Genomics (eMERGE) Network. For each individual, we mapped their ICD-9 codes to PheWAS phenotypes: cases had at least two matching codes and controls had no codes or similar diagnoses. We identified 1,572 phenotypes with at least 20 cases and performed logistic regression using the R PheWAS package adjusting for gender, study site, and the top three principal components. A Bonferroni correction of phenotypes tested (n=1572) gave a significance threshold of $3.18x10^{-5}$.

Results and Discussion

One SNP-phenotype association met our Bonferroni correction (rs425437 and lip cancer, $p=5.75x10^{-6}$) and nine additional associations had $p<5x10^{-4}$. This association is particularly interesting given the biological data gathered from our eQTL analysis. This SNP associates with increased expression of C1orf115 and MARC2 (mitochondrial amidoxime reducing component 2). While not much is known about C1orf115, MARC2 is a mitochondrial protein that has been shown to increase production of nitric oxide (NO) under hypoxic conditions.[1] Interestingly, rs425437 is predicted to have increased binding affinity for the Hypoxia-Inducible Factor 1 (HIF-1) transcription factor, possibly explaining the increased expression of MARC2. In the pathway analyses, this SNP associated with differential expression of the Huntington's Disease (HD) KEGG Pathway. Both HD and cancer of the lip have been shown to have increased NO production as part of the pathophysiology. Additionally studies have shown that HIF-1 is overexpressed in cancers of the lower lip, but not oral and larynx squamous cell carcinomas.[2] Given these data we have provided a plausible mechanism of association from SNP to transcription factor binding, increased gene expression, alteration of a molecular pathway and ultimately a disease phenotype.

References

1. Sparacino-Watkins CE, Tejero J, Sun B, Gauthier MC, et al. Nitrite reductase and nitric-oxide synthase activity of the mitochondrial molybdopterin enzymes marc1 and marc2. J Biol Chem 2014, Apr 11;289(15):10345-58.

2. Kyzas PA, Stefanou D, Batistatou A, Agnantis NJ. Hypoxia-induced tumor angiogenic pathway in head and neck cancer: An in vivo study. Cancer Lett 2005, Jul 28;225(2):297-304.

An Inter-Patient Heterogeneity Model for Biomarker Discovery Studies

Michelle Winerip, BA, Jie Wang, BS, Ji Qiu, PhD, Joshua LaBaer, MD, PhD, Garrick Wallstrom, PhD
Center for Personalized Diagnostics, The Biodesign Institute
Arizona State University, Tempe, AZ

Abstract
In this study, we motivate and present a mixture model for heterogeneous biomarker data. We discuss estimation of the model, application of the model to real biomarker data, and utility of the model for planning biomarker studies.

Introduction and Background
Inter-patient heterogeneity is a challenge in biomarker discovery studies. Molecular studies are increasingly revealing new subtypes of diseases. These subtypes often carry different prognoses and exhibit differential responses to therapies. This raises the possibility that different subtypes may also have different biomarkers. Both biological and disease heterogeneity are poorly understood. This heterogeneity impacts the experimental design of biomarker discovery studies and the analytical methods that are applied to omic data. Statistical heterogeneity models are needed for developing better analytic tools for heterogeneous data and guiding future biomarker discovery studies.

Methods
Mixture modeling is a natural approach for modeling heterogeneity. In our approach, a single parameter Box-Cox transformation of the data is modeled as a Gaussian mixture with at most four components and at most eleven total free parameters. The model can reflect different heterogeneity assumptions using constraints on the four components. The Box-Cox transformation facilitates modeling of non-Gaussian data while only requiring one additional parameter. We use the Estimation-Maximization (EM) algorithm to fit the model for an individual candidate marker. We then generate statistical inferences on the receiver operator characteristic (ROC) curve for the marker.

Results
We have developed a flexible model that balances the need to represent non-Gaussian and heterogeneous data with the ability to estimate small sample sizes. Simulation studies have shown that the model yields improved ROC curve estimation compared to the empirical ROC curve, especially for moderately small sample sizes. Estimated four-component mixture and empirical ROC curves for a proteomic breast cancer screen marker are shown in Figure 1.

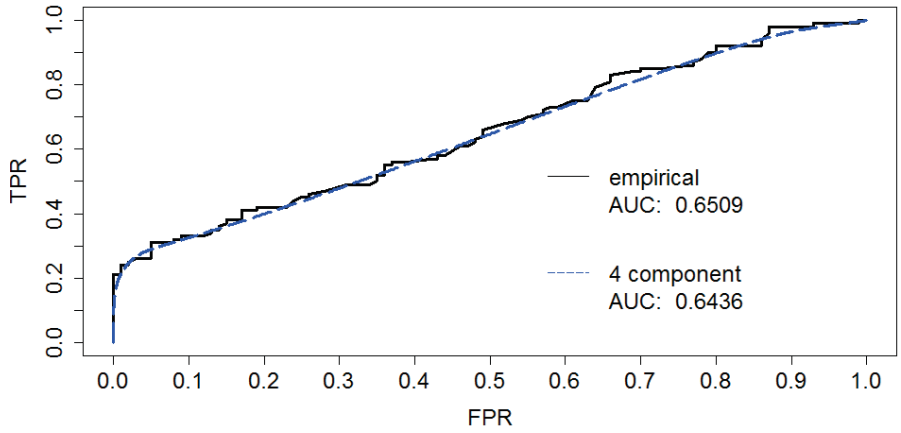


Figure 1: Empirical ROC curve (black) and a fitted four component mixture ROC curve (blue).

Conclusion
Our model can be estimated with moderately small samples sizes and yet is sufficiently flexible to fit real marker data. The model has two important applications: the development of analytic methods for heterogeneous data, such as the ROC inference described above, and applications for planning future biomarker studies. Via simulation, researchers can use the model to conduct power analyses that appropriately account for inter-patient heterogeneity.

Acknowledgement: This work was partially supported by the National Cancer Institute (R21CA187892 and 7U01CA117374).

Identification of epigenetic modifications that contribute to pathogenesis in therapy-related AML: Effective integration of genome-wide histone modification with transcriptional profiles

Xinan (Holly) Yang, Bin Wang, John M Cunningham

Section of Hematology/Oncology, Dept. of Pediatrics, Comer Children's Hospital University of Chicago

Background: Therapy-related, secondary acute myeloid leukemia (t-AML) is an increasingly frequent complication of intensive chemotherapy. This malignancy is often characterized by abnormalities of chromosome 7, including large deletions or chromosomal loss. A variety of studies suggests that decreased expression of the *EZH2* gene located at 7q36.1 is critical in disease pathogenesis. This histone methyltransferase has been implicated in repression of transcription through modification histone H3 on lysine 27 (H3k27). However, the critical target genes of EZH2 and their regulatory roles remain unclear.

Method: To characterize the subset of EZH2 target genes that might contribute to t-AML pathogenesis, we developed a novel computational analysis to integrate tissue-specific histone modifications and genome-wide transcriptional regulation. Initial integrative analysis utilized a novel "seq2gene" strategy to largely explore target genes of chromatin immuneprecipitation sequencing (ChIP-seq) enriched regions. By combining seq2gene with our Phenotype-Genotype-Network (PGnet) algorithm, we profoundly enriched genes with similar expression profile and genomic or functional characteristics into "biomodules".

Results: Initial studies using data derived from both normal and leukemic cell lines as well as murine cells deficient in EZH2 identified *SEMA3A* (semaphoring 3A) as a novel oncogenic candidate that is regulated by EZH2-silencing. A microsatellite marker at the *SEMA3A* promoter has been associated with chemosensitivity and radiosensitivity. Notably, our subsequent studies in primary t-AML demonstrate an expected up-regulation of *SEMA3A* that is EZH2-modulated. Furthermore, we have identified three biomodules that are co-expressed with *SEMA3A* and up-regulated in t-AML, one of which consists of previously characterized EZH2-repressed gene targets. The other two biomodules include MAPK8 and TATA box targets. Together, our studies suggest an important role for EZH2 targets in t-AML pathogenesis that warrant further study. Conclusion: Developed computational algorithms and systems biology strategies will enhance the knowledge discovery and hypothesis-driven analysis of multiple next generation sequencing data, for t-AML and other complex diseases.

Predicting Papillary Thyroid Carcinoma Patient Outcomes through Gene Expression Data

Kun-Hsing Yu, MD[1,2,3], Wei Wang, MBBS, PhD[4], Chung-Yu Wang, MS[3], Michael Snyder, PhD[2]

[1]Biomedical Informatics Program; [2]Department of Genetics; [3]Department of Computer Science; [4]Department of Health Research and Policy, Stanford University, Stanford, CA

Summary

Leveraging publicly available RNA-sequencing data, we predicted the survival outcomes of papillary thyroid carcinoma patients with area under receiver operating characteristic curve (AUC) of 0.9545, and identified important biological processes associated with tumor development. This informatics pipeline provided novel insights into thyroid cancer biology and contributed to personalizing cancer management.

Abstract

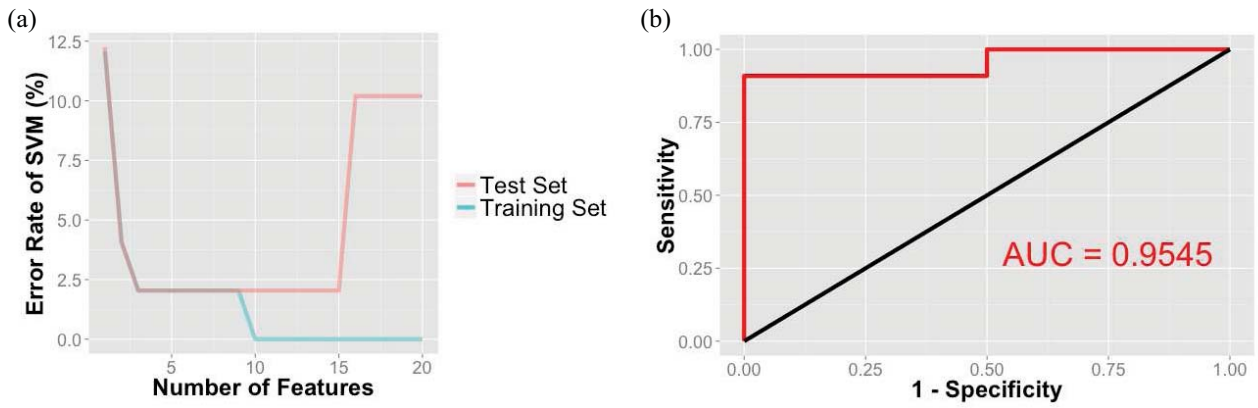
Introduction and Background: *Unlike most cancers, thyroid cancer has an increasing incidence rate over recent years, which cannot be completely accounted for by improved disease detection. Understanding the genetic and molecular basis of the disease may help us stratify patients with different outcomes and provide personalized treatment plans.*

Methods: *We acquired clinical and RNA-sequencing data of papillary thyroid carcinoma patients (n=484) from The Cancer Genome Atlas, divided them into distinct training and test sets (70% cases in the training set, and 30% cases in the test set), and dichotomized patients with known survival time into two outcome groups according to their median survival time (2000 days). We leveraged supervised machine learning methods to predict stages and survival outcomes, selected the top features by forward feature selection and Wilcoxon test statistics, and evaluated prediction performance through the independent test set. We also utilized unsupervised machine learning methods to gather biological insights into the gene expression patterns of our tumor samples.*

Results: *Using support vector machine (SVM) with the expression levels of three genes (TERT, CCDC60, and ACADSB) selected by forward feature selection, our best classifier predicted patient survival outcomes with area under receiver operating characteristic curve (AUC) of 0.9545 on the test set (Figure 1). We also predicted tumor stage with test accuracy around 70%. Factor analysis on gene expression features further identified important biological processes related to stage progression, including RNA binding, metabolism, and extracellular matrix structure.*

Discussion: *In summary, we identified novel genomic markers indicative of patient outcomes and discovered the nuanced changes in gene expression profiles as tumors progressed in stage. With the increasing availability of clinical cancer genomics data, we envision personalizing treatments based on the predicted disease outcomes, thereby increasing the quality of care and reducing the cost of thyroid cancer management.*

Figure 1. (a) Survival classification error rates of the best performing classifier (SVM with Gaussian kernel) with varying numbers of features selected by forward feature selection (b) Receiver operating characteristics (ROC) curve of survival outcome prediction of the test cases.



Prediction of Drug-Drug Interactions Based on Clinical Side Effects

Ping Zhang, PhD, Fei Wang, PhD, Jianying Hu, PhD
Healthcare Analytics Research, IBM T.J. Watson Research Center, New York, USA

Abstract

Drug-Drug Interaction (DDI) is an important topic for public health. We propose a new method to predict DDIs based on clinical side effects (SEs). We show that SEs are more predictive features than chemical structures in DDI prediction. We also find from feature selection that some SEs might result in unsafe co-prescriptions and some SEs might contribute to therapeutic effects of polypill therapy.

Introduction

DDIs may happen unexpectedly when more than one drugs are co-prescribed, causing serious side effects. Discovering and predicting DDIs will not only prevent life-threatening consequence in clinical practice, but also prompt safe drug co-prescriptions for better treatments. In this study, we propose a new method to predict adverse DDIs based on the hypothesis that clinical SEs provide a human phenotypic profile and can be translated into the development of computational models of DDIs.

Methodology

By merging drugs from SIDER (an SE database of drugs containing information on market medicines and their recorded adverse drug reactions), OFFSIDES (an SE dataset built by mining FDA AERS system while controlling confounding factors), and PubChem, we obtained 569 drugs. For each drug, it has been described as 4,192 SIDER SE features, 10,093 OFFSIDES SE features, and 881 PubChem chemical substructure features. Thus, a drug pair can be represented as a vector of features with value of 0, 1 and 2 depending whether zero, one or both drugs have such feature (i.e., SIDER/OFFSIDES SE keyword, or PubChem substructure).

Among the 569 drugs, we identified a set of 2,576 unsafe drug pairs from DrugBank, and used it as the positive training set of the DDI prediction model. We also identified a set of 349 approved pairwise drug-drug co-prescriptions from either drug combination database (DCDB) or FDA Orange Book, and used it as the negative training set of the DDI prediction model. In order to avoid overfitting, we applied L1-regularized logistic regression model to evaluate the performance.

Results and Discussion

Figure 1 shows the averaged ROC curves of 50 runs of 10-fold cross-validation for DDI prediction based on different information sources. This figure shows that SEs are more predictive features than chemical structures in DDI prediction. By using Fisher's exact test ($p<0.01$), we obtained 118 efficacy-related SEs and 897 undesired SEs. In a drug pair, when both drugs have one of the undesired SEs, the co-prescription could be unsafe and result in adverse DDIs; when both drugs only have efficacy-related SEs, the co-prescription might be safe and certain SEs would contribute to the therapeutic effects. Figure 2 provides a graphical illustration of the uses of the selected SEs to rule out unsafe co-prescriptions and to suggest polypill therapy.

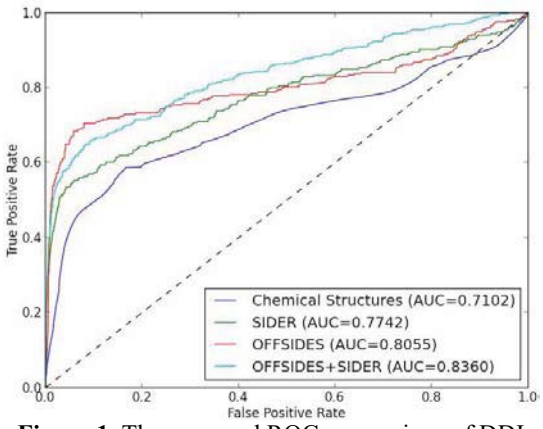


Figure 1. The averaged ROC comparison of DDI predictions based on four different information sources.

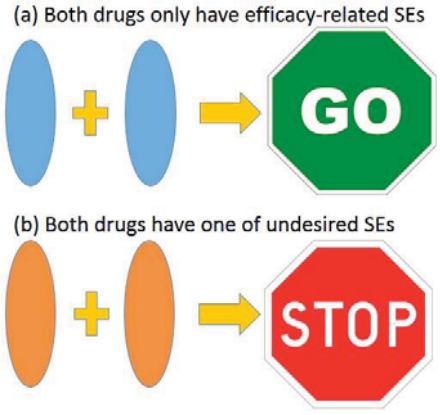


Figure 2. A graphical illustration of the uses of selected SEs: (a) suggest polypill therapy; (b) rule out unsafe co-prescriptions.

Identifying Cancer Driver Mutations and Genes from Next-generation Sequencing Data

Zhongming Zhao, PhD[1,2,3], Peilin Jia, PhD[1], Junfeng Xia, PhD[1]
[1]Department of Biomedical Informatics, [2]Department of Cancer Biology, [3]Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN, USA

Background: Recent large-scale cancer genome studies such as those from The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have led to the identification of a large number of genomic datasets (somatic mutations, gene expression, methylation, microRNA expression, etc.) in more than 20 cancer cell types through next-generation sequencing (NGS) or microarray technologies. Among those somatic mutations identified, most are passenger mutations, which do not contribute much to carcinogenesis. However, a relatively small number of mutations have been documented as driver mutations, and they confer the advantage of cell growth. To date, more than a hundred driver mutations have been uncovered. However, how to identify them, either novel ones in any cancer type or known ones but have not been reported in a specific cancer, has been a challenging task.

Methods: Through our numerous cancer genomics projects, we applied three strategies: 1) to develop pipelines to prioritize cancer driver mutations (alternatively, actionable mutations) from numerous passenger mutations in cancer genomes, 2) through systematic examination of somatic mutations in tumor genomes harboring known driver mutations versus those without having any known driver mutations (i.e. "pan-negative" tumor genomes), and 3) to develop novel computational algorithms and tools for detecting driver genes through cancer genome data.

Results: First, we developed a pipeline to prioritize numerous somatic mutations in a cancer genome by considering three features: those novel nonsynonymous somatic mutations that have high coverage and quality score from variant callers and are verified by independent experiments, those somatic mutations occur in important biological pathways (e.g. signaling pathways), and those novel mutations have occurred in our genotyping screening of tumor bank. We applied this pipeline to our melanoma whole genome sequencing (WGS) project and identified critical BRAF L597 mutations that are associated with MEK inhibitor sensitivity. This mutation site is 3-amino acid away from the well-known BRAF V600 site. Our study demonstrates the utility of WGS in identifying actionable mutations in tumors and the potential therapeutic implications of BRAF L597 mutations in melanoma. Second, we examined mutations from recently published melanoma NGS data involving 241 paired tumor-normal samples to identify potentially clinically relevant mutations. We found candidate genes whose mutations in melanoma genomes are more likely associated with the known *BRAF* or *NRAS* driver mutations, and novel genes in the "pan-negative" melanoma genomes[1]. Finally, through comparing mutation patterns (e.g. mutation clustering or hotspots) and adjustment of possible mutation biases (e.g. gene length bias), we developed two methods: 1) a personalized mutation network method, VarWalker[2], and 2) a mutation set enrichment analysis (MSEA)[3] for detecting putative cancer (driver) genes identified from NGS data. Our evaluation using the TCGA and other large-scale cancer genome datasets demonstrated both tools are promising for identifying cancer genes. Specifically to TCGA data, somatic mutations were retrieved and classified into functional groups (e.g. nonsynonymous mutations or protein domains), and then tested using our methods.

Summary: Our pipeline for analyzing somatic mutations from WGS data has been shown effective on prioritizing actionable mutations in tumor genomes. In our real application to a melanoma genome, we identified a clinically important mutation, BRAF L597, which has been included in our routine cancer gene screening. Our meta-analysis of 241 melanoma genomes provided a roadmap for the study of genes with potential clinical relevance. Finally, our computational algorithms, via both network (VarWalker) and mutation hotspots (MESA), have been found useful for identifying cancer genes in numerous cancer NGS datasets.

References:

1. Xia, J., Jia, P., Hutchinson, K. E., Dahlman, K. B., Johnson, D., Sosman, J., Pao, W. & Zhao, Z. A meta-analysis of somatic mutations from next generation sequencing of 241 melanomas: a road map for the study of genes with potential clinical relevance. Mol Cancer Ther 13, 1918-1928 (2014).
2. Jia, P. & Zhao, Z. VarWalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. PLoS Comput Biol 10, e1003460 (2014).
3. Jia, P., Wang, Q., Chen, Q., Hutchinson, KE., Pao, W. & Zhao, Z. MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis.. Genome Biol 15, 489 (2014).

SPIRIT-NLP: A Natural Language Processing (NLP) Platform for Computable Protocol

Weizhong Zhu, Ph.D. and Ajay Shah, Ph.D. City of Hope, Duarte, CA

Summary

Making eligibility criteria (EC) in clinical protocol narratives computable still remains a big challenge. SPIRIT-NLP platform is proposed to convert unstructured EC into coded and searchable forms. This system integrates open-source NLP software and provides advanced capabilities on semantic annotation, search and visual analytics.

Introduction

Lack of data standardization and non-computable protocols make it difficult to match EMR records with EC for cohort identification. Protocol narratives do not adhere to standard terminologies and ontologies. They contain abbreviations, interpretive, temporal and prospective criteria. Automated extraction of EC from protocol narratives and transforming them into meaningful representations is very complicated. We propose to utilize Natural Language Processing (NLP) to identify common and disease-specific eligibility or ineligibility. The extracted criteria from unstructured protocol text are standardized using UMLS based ontologies.

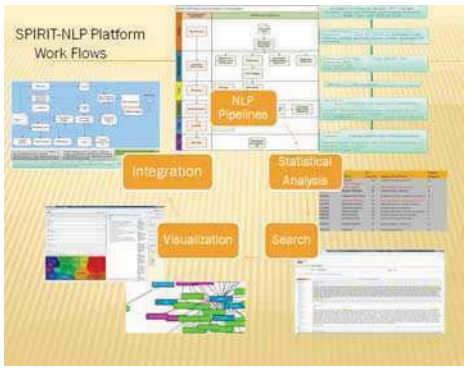


Figure 1: SPRIT-NLP with its applications

Methods

SPIRIT-NLP framework (see Figure 1), a search and knowledge discovery engine for clinical text, is designed to process protocol narratives more thoroughly. SPIRIT-NLP integrates state-of-art open-source systems under the UMLS ontology-supported NLP architecture to annotate, categorize and analyze the criteria with standard domain ontology concepts, classes, attributes and values. The system initially segments an EC as tokens and chunks. Then, a concept fusion annotator combines MetaMap and cTAKES to map these tokens/chunks to unified UMLS Concept Unique Identifiers (CUIs) belonging to semantic types under categories such as Disease, Medication, Physiology, Anatomy, Symptom, Lab Procedures and Temporal Entities. The annotator also extracts semantic information such as acronym pairs (abbreviation and expansion description) and negation triggers. ReVerb identifies phrase triplets. Based on syntactical patterns between noun chunks and verb chunks, concept coupling pairs UMLS concepts and then retrieves semantic relations between the concepts from UMLS and SemRep. These core concepts are then ranked based on the frequency of the documents they appear in and are validated in consultation with clinicians. To a core concept, measure/value/unit/triplet extractor collects the contextual information of attributes/conditions that co-occur with the concept using normalization algorithms, regular expression and statistical analysis. Then these terms, phrases, concepts, measures, values and triplets are formulated as EC rules. Finally these core concepts with their relationships are inverted and RDF indexed for keyword, concept or RDF query search.

Results

SPIRIT-NLP is applied to automatically annotate, classify, search and group the textual fields of the protocols on Clinical Trials On-Line (CTOL) system at City of Hope and the EC of active cancer-related protocols at ClinicalTrials.Gov. Solr, an open-source enterprise search engine, makes these coded protocols searchable. Hybrid schemas of Solr indexing are developed to sign weights to fields, keywords and concepts regarding to their importance and rank relevance between user queries and protocols. This search engine has fundamental capabilities such as phrase/concept based auto-complete, spelling checking, MoreLikeThis, hierarchal facet filtering by concept classes or types of Metadata, distributed query and local clustering of response pages. A case study involved a clinician to evaluate the ranking list of EC rules extracted from 388 active melanoma protocols at ClinicalTrials.Gov. The core EC identified by the clinician is also ranked in the top list of our system. Integrated with visual analytic software such as text clustering bench Carrot2 and Cytoscape, this platform demonstrates the potential for automatic classification of groups of EC and semantic network analysis.

Discussion

Combined with EMR SPIRIT-NLP will be able to automate and accelerate cohort identification. The next goal is to dynamically generate the protocol-specific eligibility nodes in SPIRIT-DT (Decision Trees module) with coded EC and automatically formulate I2B2 queries to search EMR and find cohorts of patients.

Rule-Based Extraction of Lung Cancer Stages from Free-Text Clinical Notes

Ryan Aldrich, MS; Emily Silgard, MS
Fred Hutchinson Cancer Research Center, Seattle, WA

Abstract

We developed a rule-based, natural language processing system to extract the prognostic stage of lung cancer patients at the time of initial diagnosis from the free text of clinical notes. This program was used to identify the records of stage IV patients treated at the Seattle Cancer Care Consortium for priority inclusion in outcomes research.

Introduction

Human abstractors are often called upon to read massive quantities of clinical notes to identify certain patient populations, perhaps for inclusion in research or clinical trials. This system is designed to support human abstractors by automatically identifying lung cancer patients who were diagnosed at prognostic stage IV

At diagnosis, lung cancer is assigned a stage from I to IV, but this system is optimized to function as a binary classifier that labels patients as either "stage IV" or "not stage IV". Under AJCC guidelines, the presence of metastatic disease at diagnosis is the defining characteristic of stage IV. Thus, this system can been seen as discriminating patients who were diagnosed with metastatic lung cancer from those who were diagnosed with more localized lung cancer and later developed metastatic disease.

Prior work by Nguyen[1] has shown that rule-based approaches are effective in extracting metastasis staging information from the free text of pathology reports.

Methods

The dataset used to develop this system was a corpus of 21,535 clinic notes regarding 485 lung cancer patients, for which human abstractors had labeled each patient's diagnosed prognostic stage. In this dataset 206 (42%) patients were labeled as "stage IV", while the remaining 279 (58%) were other stages. Of these patient files, 60% were used for development and 40% were held out for final evaluation. These sets were randomly assigned, but normalized on both the percentage of patients per stage and the number of clinical notes per patient file.

The system utilizes a series of rules, wherein each rule is allowed to "vote" for a prognostic stage. Rules are weighted, and may assign zero, one, or multiple votes to each clinical note. All rules are applied to all notes. A patient's votes are the sum of their notes' votes. The stage with the most votes is then assigned as the predicted stage.

The most productive rule captures 98% of the true positive cases simply by looking for the terms "metastatic", "palliative", or "stage IV". A negation detection algorithm is used to help eliminate some false positives (e.g. "no evidence of metastasis"). Other rules prevent false positives by accounting for a variety of factors: term disambiguation (e.g. "stage IV kidney disease" vs "stage IV lung cancer"), hypothetical and hedging language, semantic role errors (e.g. "the patient's brother had stage IV lung cancer"), and patients who developed metastatic disease after diagnosis.

Results

On the final evaluation set the system had an accuracy of 94%, recall of 96%, and precision of 90%. The accuracy results are the same score as reported by Nguyen[1] for determining M staging. However our system achieves this without depending on a computationally expensive external thesaurus.

Discussion

Future work will include expanding the framework and logic of the existing system to include the classification of all four prognostic stages as well as extending the system to work with additional types of cancer.

References

1. Nguyen AN, Lawley MJ, Hansen DP, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc [Internet]. 2010 [cited 2014 Jun 23];17(4):440–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995652/

Title: Pharmacomicrobiomics of Irritable Bowel Disease

Tomer Altman, David A. Relman, and David L. Dill
Stanford University

The human gut microbiome teems with a dizzying array of microbial flora. Recent studies have correlated microbial diversity with clinical phenotypes such as obesity, irritable bowel syndrome, and even cardiovascular disease. Despite playing important roles in immune response, nutrition, metabolism of pharmaceuticals, and enteric diseases, little is known of the intricate network of metabolic transformations mediated by the microbiota. In particular, the vast variety of microbial enzymes are able to metabolize pharmaceuticals, yet little is known about the type and degree of interaction.

Using the GutCyc computational model of the biochemical network present in the gut ecosystem, we find pharmaceuticals with oral or anal routes that have biochemical interactions. Using next-generation sequencing data from the MetaHIT study on irritable bowel disease, we mine the relative abundance of the interacting enzymes and search for correlation with clinical phenotypes such as Crohn's Disease or ulcerative colitis.

Finally, using correlated taxonomic markers, we propose diagnostic tests that will allow clinicians to predict whether a patient is likely to have significant interaction between their gut microbiome and the administered drug. This has immediate clinical implications, in that it may guide clinicians to choose alternate drugs or drug routes based on a patient's microbiome. This will provide a first step in establishing pharmacomicrobiomics as a nascent specialty within personalized medicine.

Funding for this work comes from the NIH Biotechnology Training Grant, KAUST, and the Stanford School of Medicine.

Widespread parainflammation in human cancer

Dvir Aran[1], Audrey Lasry[2], Adar Zinger[2], Moshe Biton[2], Eli Pikarsky[2], Yinon Ben-Neriah[2] and Asaf Hellman[1]

[1] Developmental Biology and Cancer Research and [2] The Lautenberg Center for Immunology and Cancer Research, IMRIC, Hebrew University—Hadassah Medical School

Background: Chronic inflammation is recognized as an enabling hallmark of cancer, but the mechanisms by which inflammation promote most cancers remain obscure. We recently showed that parainflammation - a unique variant of inflammation, under-recognized due to absence of immune cell infiltrate, can strongly promote gut tumorigenesis in the mouse upon p53 loss. Here we explored the prevalence and correlates of parainflammation in human cancer.

Results: Based on the mouse model and further refinement in human datasets, we generated a parainflammation transcriptome signature. Using expression data from the cancer genome atlas (TCGA: 6,511 cancer and 635 normal samples) and the cancer cell line encyclopedia (814 samples) we examined the occurrence of parainflammation in human cancer, its association with common cancer mutations and its relevance to cancer prognosis. Epithelial organoids generated from tumor and normal intestinal tissues of different mouse models were used to validate the refined PI signature.

We identified a parainflammation gene signature in 27% of all human tumor samples and 23.8% of carcinoma cell lines, derived from the majority of TCGA tumor types, irrespective of canonical inflammation. We revealed a tight association between the presence of parainflammation and p53 mutations as well as with worse prognosis.

Conclusion: Parainflammation, a low-grade inflammation form, originating autonomously in cancer cells, is widely prevalent in human cancer. It characterizes half of all human cancer types, particularly those commonly harboring p53 mutations. Our data suggests that parainflammation may be a major driver for p53 mutagenesis and a guide for cancer prevention by NSAID treatment.

G-DOC *Plus*: The next-generation systems medicine platform for precision medicine

Krithika Bhuvaneshwar, MS[1], Anas Belouali, MS[1], Varun Singh, MS[1], Robert M Johnson, *MS*[1], Lei Song, MS[1], Adil Alaoui, MS[1], Shruti Rao, MS[1], Michael Harris, MA[1], Yuriy Gusev, PhD[1], Subha Madhavan, PhD[1]
[1]Innovation Center for Biomedical Informatics, Georgetown University, Washington DC

Summary

G-DOC Plus is an enhanced web platform that uses cloud computing and other advanced computational tools to handle next generation sequencing (NGS) and medical images so that they can be analyzed in the full context of other omics and clinical information. G-DOC Plus tools have been leveraged in the cancer and non-cancer realms for hypothesis generation in precision medicine and translational research.

Introduction and Background

Our flagship web platform, the Georgetown Database of Cancer (G-DOC), was deployed in April 2011 to enable the practice of an integrative translational and systems-based approach to research and medicine in cancer. G-DOC is a feature-rich shareable research infrastructure that allows physician scientists and translational researchers to mine and analyze a variety of "omics" data in the context of consistently defined clinical outcomes data for cancer patients.

With the explosion of next generation sequencing (NGS) in 2007, the size and complexity of genomic data has increased many fold since then, making its analysis, management and integration increasingly challenging. Scientists today are using not only a combination of clinical, NGS and omics data for analysis, but also medical and pathology images for validation of analysis results. To drive hypothesis generation and validation of molecular markers for translational research, it is convenient to have a "one–stop" system that can handle all these data types in one location. For this purpose, we expanded the G-DOC system to support NGS and medical images. Moreover, the success of G-DOC in the cancer realm has helped us realize the importance of such systems in non-cancer, for complex diseases such as Alzheimer's, Duchene Muscular dystrophy, and others.

With the goal of improving overall health outcomes through advanced genomics research, we present G-DOC *Plus*, our web platform that enables the integrative analysis of multiple data types to understand mechanisms of cancer and non-cancer diseases for precision medicine. G-DOC *Plus* allows researchers to explore data one sample at a time, as a sub-cohort of samples; or as a population as a whole, providing the user with a comprehensive view of the data.

Methods

G-DOC *Plus* was developed using our in-house architectural framework to support over 600 users and 9000 patient and cell line data. Our data collection includes whole genome sequencing (WGS) data from the 1000 Genomes Project and Complete Genomics; multi-omics data from the NCI-60 data collection; numerous breast, GI, and pediatric cancer studies; and non-cancer studies including Duchenne Muscular Dystrophy (DMD) and Alzheimer's disease. G-DOC *Plus* includes a broad collection of bioinformatics and systems biology tools for analysis and visualization of many 'omics' types including DNA, mRNA, microRNA, metabolites, copy number and WGS data in addition to clinical data such as demographics, pathology, and outcomes. Users can also explore somatic mutations and perform gene enrichment using Reactome. Processed NGS data and medical images are integrated within this web portal to enable their analysis in the full context of other omics and clinical information.

Results

G-DOC *Plus* tools have been leveraged in cancer and non-cancer for hypothesis generation; multi-omic in-silico and population genetics analysis. It has also been used to support detection of prognostic markers for relapse in colorectal cancer samples, understand molecular changes linked with pediatric cancer, detect key metabolites related to disease severity and to examine breast cancer MRI images in conjunction with clinical data.

Conclusion

We present G-DOC *Plus*, with a long-term vision of extending this systems medicine platform to hospital networks to provide more effective clinical decision support using multi-omics and NGS data. G-DOC *Plus* was released in October 2014, and is available at: https://gdoc.georgetown.edu.

Using Electronic Health Records to Uncover Disease-Birth Month Dependencies

Mary Regina Boland, MA[1-2], Zachary Shahn, MS[4], David Madigan, PhD[4], George Hripcsak, MD, MS[1], Nicholas P Tatonetti, PhD[1-3]

[1]Department of Biomedical Informatics, [2]Department of Systems Biology, [3]Department of Medicine, [4]Department of Statistics, Columbia University

Summary

Many studies investigate relationships between early developmental seasonal factors and disease-risk using birth month as a proxy. These prior methods suffer from disease selection bias by choosing to investigate popular over rare diseases. They also exhibit publication bias, as only studies finding an association are likely to be published. To address these biases, we developed a high-throughput hypothesis-free method that investigates every condition's association with birth month. We use Electronic Health Records from Columbia University Medical Center. We found 55 significant disease-birth month associations including 19 explicitly reported in the literature, 20 closely related to previously reported associations and 16 novel associations ($p<0.001$). Nine of these novel associations were for cardiovascular conditions. Overall, patients born in May and July had the lowest disease risk. Patients born in October and November had the highest disease risk.

Introduction and Background: Hippocrates noted in 460 BCE, *"for with the seasons the digestive organs of men undergo a change*" in reference to a connection he observed between seasonality, early development, and disease. More recently, studies have linked birth month with a number of diseases including neurological[1], endocrine[2], inflammatory[3] and reproductive conditions[4]. These hypothesis-driven analyses suffer from selection bias in that they typically favor popular over rare diseases. In addition, there is publication bias as negative findings are rarely published[5]. Here, we present a hypothesis-free approach to investigate birth month associations with 1,688 diseases using the Electronic Health Records (EHR).

Methods: We conducted a retrospective epidemiological study. Our study population consisted of 1,749,400 patients born between 1900-2000 who have records at New York-Presbyterian/Columbia University Medical Center and was conducted with Institutional Review Board approval. We investigated associations with birth month across all recorded conditions. A condition is defined as any *Systemized Nomenclature for Medicine-Clinical Terms* code with at least 1,000 patients (1,688 total 8conditions). Associations between birth month and disease were modeled as a logistic regression and significance assessed using a chi-square test (R version 3.1.0). We randomly sampled 10 control patients without the condition for each case patient. We report all associations with p-values passing false discovery rate correction[6]. For each associated condition, we report the birth month conferring the greatest risk along with the birth month conferring the greatest protection.

Results: We reviewed 92 relevant PubMed-indexed articles on birth month and curated a reference set of associated conditions. We evaluated our results against this reference set (PPV=0.345; $p<0.001$). We found 55 significant disease-birth month associations, including 19 that were explicitly reported previously in the literature, 20 closely related to previously reported associations, and 16 previously unreported associations. Nine of these novel associations were for cardiovascular conditions and another was for malignant prostate cancer (greatest risk among those born in March). Overall, patients born in May and July had the lowest disease risk (May: 0 at risk and 3 protective conditions; July: 0 at risk and 6 protective conditions). Patients born in October and November had the highest disease risk (Oct: 15 at risk and 7 protective conditions; Nov. 15 at risk and 3 protective conditions).

Discussion: Our study revealed nine cardiovascular conditions associated with birth month (e.g., atrial fibrillation). We found the greatest cardiovascular disease risk for those born in February. Atrial fibrillation is associated with birth weight [7], which depends on birth month,[8] suggesting a rationale for the association with atrial fibrillation. In addition, children born to survivors of H1N1 1918 subtype have been associated with a >20% excess risk of cardiovascular disease,[9] suggesting a relationship between maternal infection and cardiovascular disease risk that is independent of maternal malnutrition[9].

We also found an association between malignant prostate cancer and birth month. Maternal estrogen expression levels, which exhibit seasonal variation,[10] and maternal birth month dependency[4] are known to affect fetal prostate development prenatally[11]. Estrogen also plays a role in preventing prostate cancer growth[12]. Our findings support the hypothesis of a link between prenatal prostate development and lifetime risk of prostate cancer mediated through regulation of estrogen expression. Study limitations include the lack of condition independence (conditions rarely occur in isolation) potentially affecting multiplicity correction.

Our study finds disease-birth month associations in a high-throughput hypothesis-free manner. We confirm known disease-birth month associations and discover associations that bolster existing hypotheses.

Acknowledgments: Support for this research provided by **T15 LM00707**, **LM006910**, and **R01 GM107145**. Authors report no conflicts of interest.

References

1. Willer CJ, Dyment DA, Sadovnick AD, Rothwell PM, Murray TJ, Ebers GC. Timing of birth and risk of multiple sclerosis: population based study. BMJ. 2005;330(7483):120.
2. Kahn HS, Morgan TM, Case LD, Dabelea D, Mayer-Davis EJ, Lawrence JM, et al. Association of Type 1 Diabetes With Month of Birth Among US Youth The SEARCH for Diabetes in Youth Study. Diabetes Care. 2009;32(11):2010-5.
3. Disanto G, Chaplin G, Morahan JM, Giovannoni G, Hypponen E, Ebers GC, et al. Month of birth, vitamin D and risk of immune mediated disease: a case control study. BMC medicine. 2012;10(1):69.
4. Huber S, Fieder M, Wallner B, Moser G, Arnold W. Brief communication: birth month influences reproductive performance in contemporary women. Hum Reprod. 2004;19(5):1081-2.
5. Dickersin K. The existence of publication bias and risk factors for its occurrence. JAMA. 1990;263(10):1385-9.
6. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.
7. Conen D, Tedrow UB, Cook NR, Buring JE, Albert CM. Birth weight is a significant risk factor for incident atrial fibrillation. Circulation. 2010;122(8):764-70.
8. Chodick G, Shalev V, Goren I, Inskip PD. Seasonality in Birth Weight in Israel: New Evidence Suggests Several Global Patterns and Different Etiologies. Annals of Epidemiology. 2007;17(6):440-6.
9. Mazumder B, Almond D, Park K, Crimmins EM, Finch CE. Lingering prenatal effects of the 1918 influenza pandemic on cardiovascular disease. Journal of Developmental Origins of Health and Disease. 2010;1(01):26-34.
10. Fusani L, Van't Hof T, Hutchison JB, Gahr M. Seasonal expression of androgen receptors, estrogen receptors, and aromatase in the canary brain in relation to circulating androgens and estrogens. Journal of Neurobiology. 2000;43(3):254-68.
11. Driscoll SG, Taylor SH. Effects of prenatal maternal estrogen on the male urogenital system. Obstetrics and gynecology. 1980;56(5):537-42.
12. Härkönen PL, Mäkelä SI. Role of estrogens in development of prostate cancer. The Journal of Steroid Biochemistry and Molecular Biology. 2004;92(4):297-305.

Identifying the Impact of Genomic Variation on Glycemic Response to Metformin Using EHR-Linked Biorepository Data

Matthew K. Breitenstein, PhD[1,2], Euijung Ryu, PhD[1], Liewei Wang, MD, PhD[1], Sebastian M. Armasu, MS[1], Richard M. Weinshilboum, MD[1], Gyorgy Simon, PhD[2], Jyotishman Pathak, PhD[1]
[1]Mayo Clinic, Rochester, MN; [2]University of Minnesota, Minneapolis, MN

Abstract

Metformin is a first-line antihyperglycemic agent commonly prescribed in type 2 diabetes mellitus (T2DM), but whose pharmacogenomics are not clearly understood. In this study we seek to replicate pharmacogenetic signals and their association with glycemic response to metformin treatment using EHR-linked biorepository data.

Background and Introduction

Metformin is recommended as a first-line therapy for type 2 diabetes mellitus (T2DM)[1] and is believed to be the most prescribed drug worldwide[2]. Evidence is also accumulating that highlights the potential repurposing of metformin for cancer prevention and treatment[3]. However, the details underlying the molecular mechanism of action for metformin are not fully understood[2], particularly genetic variation in clinically relevant targets of metformin[4]. Metformin is primarily utilized to regain glycemic control in diabetic or pre-diabetic patients. Metformin is a relatively safe antidiabetic therapy[5]. However, serious adverse reactions can occur[6]. The pharmacokinetics (PK) of metformin, the transportation throughout the body, are moderately understood[7]. However, the pharmacodynamics (PD) of metformin, the physiological and biochemical impact of metformin in the body, are not clearly understood[7]. There is considerable variation in glycemic response to metformin[7]. While genetic factors may partially explain glycemic response to metformin, further studies are needed to understand the impact of variation in key transporter genes on glycemic response in clinical populations[2]. Our study aims to add clarity to metformin pharmacogenomics by understanding the impact of common variants in metformin candidate genes (n=17) on altered glycemic response in a clinical population.

Materials and Methods

Candidate genes selected for inclusion in this study are suspected metformin PK or PD determinants as designated in systematic reviews of metformin pharmacogenomics[2,5,7,8,9]. Gene-level and SNP-level analyses were performed in this study to identify genes significantly associated with change in glycemic response after exposure to metformin and directionality of the effect of corresponding SNPs. Our analysis cohort consisted of 258 T2DM patients who had new metformin exposure, existing genomic data, and longitudinal electronic health record(EHR)s. Change in glycemic response to metformin exposure via A1c measures pre and post metformin exposure serve as the outcome of interest. EHR data was utilized to develope clinical phenotypes. After quality control, gene-level and SNP-level analysis were conducted on 17 candidate genes and 463 SNPs within those genes was performed, controlling for key covariates of sex, age, and body mass index (BMI) at index metformin exposure.

Results

PRKAB2, the gene encoding the beta subunit 2 of adenosine monophosphate-activated protein kinase complex, was associated with marginally significant (p=0.0194) change in glycemic response after exposure to metformin. The next most significant association (p=0.0614) was SLC29A4, the gene encoding the plasma membrane monoamine transporter expressed in the intestine.

Conclusion

We were able to replicate one metformin PD determinate (PRKAB2), with rs7541245 having the strongest SNP association, which appeared to be marginally associated with decreased glycemic response after exposure to metformin.

References

1. Inzucchi SE, Bergenstal RM, Buse JB, et al. Management of Hyperglycemia in Type 2 Diabetes: A Patient-Centered Approach. Diabetes Care. 2012;35:1364-1379.
2. Todd JN, Florez JC. An update on the pharmacogenomics of metformin: progress, problems and potential. Pharmacogenomics. 2014;15(4):529-539.
3. Franciosi M, Lucisano G, Lapice E, Strippoli GF, Pellegrini F, Nicolucci A. Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review. PloS one. 2013;8(8):e71583.
4. Wang L, Weinshilboum R. Metformin pharmacogenomics: biomarkers to mechanisms. Diabetes. Aug 2014;63(8):2609-2610.
5. Graham G.C., Punt J, Arora M, et al. Clinical Pharmacokinetics of Metformin. Clinical Pharmacokinetics. 2011;50(2):81-98.
6. Bailey CJ, Path MRC, Turner RC. Metformin. The New England Journal of Medicine. 1996;334(9):574-579.
7. Gong L, Goswami S, Giacomini KM, Altman RB, Klein TE. Metformin pathways: pharmacokinetics and pharmacodynamics. Pharmacogenetics and genomics. Nov 2012;22(11):820-827.
8. Viollet B, Guigas B, Sanz Garcia N, Leclerc J, Foretz M, Andreelli F. Cellular and molecular mechanisms of metformin: an overview. Clinical science. Mar 2012;122(6):253-270.
9. Chen S, Zhou J, Xi M, et al. Pharmacogenetic Variation and Metformin Response. Current Drug Metabolism. 2013;14:1070-1082.

Connecting chemical structure to cellular response: an integrative analysis of gene expression, bioactivity and structural data across 11000 compounds

Bin Chen[1] (PhD), Peyton Greenside[2], Hyojung Paik[1] (PhD), Marina Sirota[1] (PhD), Atul J Butte[1] (PhD, MD)

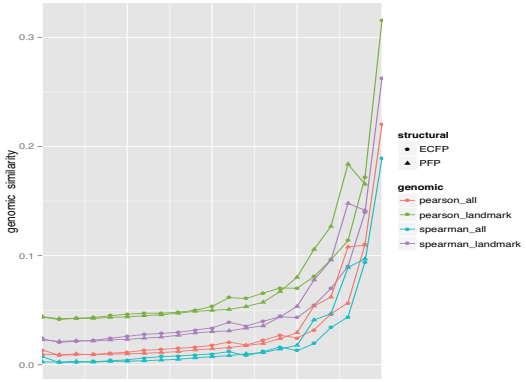
1. Division of System Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, CA
2. Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA

Summary:
Do structurally similar compounds share similar gene expression profiles? Our systematic analysis seeks to address this fundamental question by using gene expression data from LINCS and compound structure and bioactivity data from PubChem.

Background:
In translational bioinformatics and systems pharmacology, one central premise that structurally similar molecules have similar biological response is widely exploited, for example, in the creation of structurally diverse compound libraries for high-throughput screening. However, the premise does not hold sometimes. An increasing interest has evolved to measure chemical similarity using biological response data (e.g., bioactivity and phenotypic data). One critical way to assessing chemical similarity is to examine the similarity of cellular response upon compound treatment. The recent effort on the large-scale creation of the Library of Integrated Network-Based Cellular Signatures (LINCS) offers an unprecedented opportunity to quantify the correlation between chemical structure and cellular response. Gene expression is one of the most widely used cellular signatures in the characterization of cellular response.

Methods:
By integrating gene expression data from LINCS with compound structure and bioactivity data from PubChem, we perform a large-scale correlation analysis of chemical structures and gene expression profiles of over 11000 compounds taking into account cofactors such as biological conditions (e.g., cell line, dose, treatment time), compound physical properties and compound bioactivities. We first identified chemical compound pairs where two compounds were administrated under the same conditions (i.e., cell line, dose and treatment duration), and then retrieved gene expression profiles of each compound from LINCS. We retrieved chemical structures, physical properties (e.g., AlogP) and bioactivities (e.g., IC50 in a biological assay) of these compounds from PubChem. The genomic similarity and structural similarity of these pairs were assessed using several measures. Extended Connectivity Fingerprint 4 (ECFP4) and Pharmacophore Fingerprint (PFP) were used for assessing chemical similarity, and Spearman correlation and Pearson correlation were used for assessing genomic similarity. As only 1000 landmark genes were profiled by LINCS, and the expressions of the remaining genes were imputed, the expression of landmark genes and whole-genome were analyzed separately.

Results
As shown in the figure, when two compounds are structurally identical (TC = 1), the average genomic similarity using pearson correlation of landmark genes reaches a correlation value of 0.3. When structural similarity is less than 0.75 (TC <0.75), genomic similarity becomes flat, with value less than 0.1. It indicates that structural similarity does not correlate to genomic similarity when two compounds are not similar (e.g., TC < 0.75), regardless of the similarity measures. Furthermore, we show that the correlation between chemical structure and gene expression highly depends on biological conditions and biological activity. There is about a 20% chance that two structurally similar compounds (TC >0.85) share similar gene expression profiles, but regardless of structural similarity, two compounds tend to share similar profiles in a cell line when they are administrated with a higher dose or they are active in that cell line. We also observe that many structurally similar compounds only share similar genomic profiles in a right cell line with a right dose and right treatment duration.



FHIR-Based Web Services for Computational Phenotyping from Electronic Health Records

Robert Chen[1], Mark Braunstein[2], Bradley Malin[3,4], Joshua Denny[3], Abel Kho[5], Joydeep Ghosh[6], Jimeng Sun[1]

[1]School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, Atlanta, GA; [2]School of Interactive Computing, College of Computing, Georgia Institute of Technology, Atlanta, GA; [3]Department of Biomedical Informatics, School of Medicine, Vanderbilt University, Nashville, TN; [4]Department of Electrical Engineering and Computer Science, School of Engineering, Vanderbilt University, Nashville, TN; [5]Feinberg School of Medicine, Northwestern University, Chicago, IL; [6]Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX

Summary: Phenotyping from electronic health record (EHR) data can play an important role in informing clinical decision making, but current phenotyping algorithms require human supervision and are not interoperable between different data formats. We implemented a web service that computes phenotypes automatically after a user uploads EHR data following the standardized FHIR data model. Our proof of concept should motivate further efforts to develop interoperable platforms for phenotyping.

Background: Adoption of electronic health records (EHRs) by hospitals and physicians has increased dramatically since 2008 as a result of the federal government's HITECH program. [1] A major goal of Stage 3 of this program is improved clinical decision making. This is hindered partially by the difficulty of extracting meaningful concepts, or phenotypes, from EHR data. [2] Current approaches to phenotyping often involve expert specification, and can be labor intensive. Despite much research [3], a standardized platform for phenotyping using large databases of patients is still lacking. [4,5] One major obstacle is the lack of a practical and easily implementable standard clinical data model. The Fast Healthcare Interoperability Resources (FHIR) Specification [6] is gaining rapid acceptance as such a standard. [7] To facilitate interoperability, scalability and ready access to our phenotyping algorithm, we implement it as a web service that takes as input patient data represented as FHIR "resources", calculates phenotypes in the back end, and displays results via an API and web application.

Methods: EHR data is first converted to the standardized FHIR data model which is specified as a set of "resources". Three FHIR resources are used for the phenotyping algorithms: Patient, Admission, and Medication. The FHIR files are stored in a JSON format and can be constructed from EHR data such as hospital admissions, diagnoses, medication orders and general patient information. The web application allows users to upload patient data mapped to FHIR resources and to specify which phenotyping algorithms to run on the uploaded data. A phenotyping algorithm running in the back end is then applied to the uploaded patient data. Currently, we implement a set of algorithms from PheKB [8] which automatically computes phenotypes using a rule-based algorithm based upon diagnostic and medication events. Additional algorithms may be implemented as plug-ins to the back end, allowing for flexible usage. The API returns phenotyping results as another JSON file, which is displayed to the user via a web application (see Figure 1).

As a proof of concept for the web service, we built a web application hosted on Amazon Web Services, a cloud computing service, which allows users to upload patient data in the form of FHIR files and to run a phenotyping algorithm based upon user-specified parameters. The web application reports to the user a list of phenotypes computed from the uploaded dataset.

To test the system, we first converted data for a subset of 1,000 patients from the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database [9] into standardized FHIR files. We uploaded the files onto the web service, choosing to run a phenotyping algorithm which we implemented. After this step, back end algorithms are triggered by the web interface to compute phenotypes. The current implementation of the algorithms require diagnoses to be coded using the ICD-9 scheme. Further, medications are used in the uploaded format without further processing. The results of the algorithm were displayed to the user after the algorithm execution finished. Other than the initial data upload and algorithm choice step, all functions of the software are fully automated and require no human supervision.

Results and Discussion: This web application provides a convenient way for physicians to upload patient data and quickly discover patient phenotypes in the dataset. The process involves minimal human supervision apart from the data uploading step. This web application has practical significance, as it takes input data following a standardized format, addressing pressing concerns with interoperability between EHRs. In a future version we plan to improve the analytics back end and the web interface by allowing users to specify phenotyping input parameters and view phenotypes for specific patient subgroups. Furthermore, while we address the issue of structural heterogeneity by using the standardized FHIR data model, we plan to address terminological heterogeneity by aggregating data for similar events into more interpretable classes (e.g., combining similar diseases or medications into classes before calculating phenotypes). To facilitate the end goal of improved clinical decision-making, we plan to incorporate machine learning algorithms into the back end for use cases such as risk prediction for life-threatening conditions (i.e., sepsis or myocardial infarction) based upon phenotypes. While we currently focus on the usage of phenotyping in the ICU setting, the web service is scalable to large datasets and has the capacity to incorporate a wide variety of phenotyping algorithms and use cases.

Conclusion: We have built a working prototype of a computational phenotyping web service. We successfully converted an EHR dataset into standardized FHIR files, uploaded the data via an API, and ran a phenotyping algorithm on the uploaded data. Our API provides the results of the phenotyping algorithm to a web application for viewing by the end user.

Acknowledgements: This work was supported by the National Science Foundation, award #1418511. We thank Keegan Nesbitt, Sizhe Lin, Amy Zhen, and Yen Huang for assistance with the software implementation.

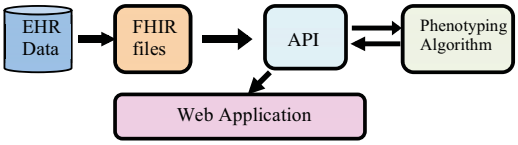


Figure 1: An illustration of the general framework for the phenotyping web service.

References:

[1] Hsiao CJ and Hing E. Use and Characteristics of Electronic Health Record Systems Among Office-based Physician Practices: United States, 2001–2013. NCHS Data Brief, No. 143; 2014.

[2] Pathak, J., Kho, A. N. & Denny, J. C. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. J. Am. Med. Inform. Assoc 2013; 20(e2):206–e211.

[3] McCarty CA, Chisholm RL, Chute CG, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011; 4(13).

[4] Pathak J, Bailey KR, Beebe CE, et al. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. J Am Med Inform Assoc 2013;20(e2):e341–e348.

[5] Rea S, Pathak J, Savova G, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: The SHARPn project. J Biomed Inform 2012;45(4):763–71.

[6] FHIR Specification Home Page;2014. [cited 2014 Sept 25]; Available from: http://www.hl7.org/implement/standards/fhir

[7] Braunstein M. Free the health data: Grahame Grieve on FHIR. InformationWeek Healthcare. [cited 2015 Jan 7]; Available from: http://www.informationweek.com/healthcare/electronic-health-records/free-the-health-data-grahame-grieve-on-fhir/a/d-id/1297761

[8] PheKB. Vanderbilt University; 2012. [cited 2013 Mar 13]; Available from: http://phekb.org.

[9] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II: A public-access intensive care unit database. Crit Care Med 2011;39(5):952-60.

SysBioCube tools for Integration of Omics and Phenotype Datasets

Sudhir Chowbina[1], Raina Kumar[1], Rasha Hammamieh[2], Marti Jett[2], Uma Mudunuri[1]
[1]Advanced Biomedical Computing Center, Frederick National Laboratory for Cancer Research, Frederick, MD
[2]US Army Center for Environmental Health Research (USACEHR), Frederick, MD

Introduction

SysBioCube is an integrated data warehouse and analysis platform for trauma-related conditions and diseases of military relevance. The system was developed by the Frederick National Laboratory for Cancer Research (FNLCR) for the US Army Medical Research and Materiel Command Systems Biology Enterprise (USMRMC-SBE). It provides browsing, querying and visualization capabilities for better understanding of the systems biology of PTSD through the integrated environment.

Background

Integrated approaches that combine genome, transcriptome, proteome, epigenome and metabolome profiling have become important since they provide better understanding of the biological systems. An effective way to interpret complex omics experiments is the combined visualization of the experimental data with existing knowledge and phenotypic data. Networks, charts and heatmaps are effective tools to support human reasoning, and when different data types are presented in the context of the biological pathways, much information can be gained about the biological mechanisms.

Results

SysBioCube supports multiple experimental data types and multiple projects. The web interface offers many options to find biological associations from different experimental results. Three such options are described here: (a) Integrome is a SysBioCube application which allows browsing of data sets and its associations based on user-specified criteria. The main page displays a web form where the given data files can be chosen according to available Diseases, Species, Experiment Types, Strains (in case of experimental animals), Groups and Tissue. Based on the selection, Integrome computes automatic association across experiments, within experiments and across annotation databases (e.g., Gene-Drug, Gene-Disease) to display the information in table and network format. (b) Heatmap provides omics to clinical data associations through 2D displays of the values in a data matrix. It represents the level of expression of many entities (e.g., genes) across a number of comparable samples. The users can see the cluster of entities with similar expression, while simultaneously visualizing the normalized entity expression values alongside the clinical variables. (c) The correlation module provides binary association between log fold-change of entities of any two given experiments. This enables users to visually explore fold-change directions and ascertain if the changes are as expected under normal conditions.

Conclusion

Here, we present tools developed with SysBioCube to derive association within experiments, across experiments and across annotation databases enriched with experimental data, and also to visualize associations as a network or heatmap. SysBioCube is an integrative knowledge base that will help investigators access, visualize and analyze comprehensive information about military-relevant diseases.

Reference: Sudhir Chowbina, Rasha Hammamieh, Raina Kumar, Nabarun Chakraborty, Ruoting Yang, Uma Mudunuri, Marti Jett, Joseph M Palma, Robert Stephens. SysBioCube: A Data Warehouse and Integrative Data Analysis Platform Facilitating Systems Biology Studies of Disorders of Military Relevance. *AMIA Jt Summits Transl Sci Proc* 2013 18;2013:34-8. Epub 2013 Mar 18.

Functional characterization of disease-associated genes in Autism Spectrum Disorders using intelligent literature extraction.

Maude M. David[1], Alp Ozturk[1], Vanessa V. Sochat[1], Jae-Yoon Jung[1], Dennis P. Wall[1]

[1]Department of Pediatrics, Division of Systems Medicine, Stanford University, Stanford, CA 94305

Summary: Using a rule-based text-mining algorithm (Jung et al., 2014), SFARI genes and phenopedia (Yu et al., 2010), we identified 1006 genes involved in Autism Spectrum Disorder (ASD), and 4595 genes associated with comorbid disorders of ASD. We conducted a cross disorder analysis by identifying the pathways involving these genes, and targeting 58 pathways showing a differential expression profile in autistic patients, especially pathways related to immune functions.

Background: Autism Spectrum Disorder (ASD) is a heterogeneous and heritable developmental disorder that affects 1 in 68 children. ASDs are not one but many conditions and finding relevant biomarkers for this condition may be more related to the root of narrow sense autism rather than those genes on the boarders that do overlap with other conditions. The scientific community has exhaustively studied this disorder, and the literature about the potential genes involved in this condition is massive. A systematic screening of all genes already highlighted by the literature would allow for the identification of the most promising that could be tested for diagnostic validity.

Methods: To perform a complete and systematic screening of the ASD literature and identify genes potentially involved in autism, we leverage robust methods in text mining (Jung et al., 2014, Yu et al., 2010), and trusted standards for gene set enrichment to explore the similarities and differences between autism and related conditions. Using the KEGG API and KEGG BRITE database, we characterized the biological functions of these genes and mapped them to corresponding KEGG pathways. Each pathway was subsequently tested for its potential enrichment in expression data available in the Gene Expression Omnibus (GEO) database.

Results: We identified 1006 genes that are positively correlated to ASD in the literature. 347 of these genes were unique to autism, and the remaining 3936 had associations with 35 comorbid disorders. Functional mapping revealed 153 pathways, including 58 that showed significant differential expression between autistic patients and control. Among these pathways, several were related to the immune system, and more specifically, to the reaction toward bacterial and viral infection. **Discussion:** In this study we parsed the high level of autism genetic complexity by conducting cross disorder analysis, and identify several biomarkers to autism. The next step to validate these biomarkers will be to integrate them with exome sequencing to identify the variants associated with them, and to estimate their diagnostic validity via development of a machine learning classifier.

Planning clinically relevant biomarker validation studies using the "number needed to treat" concept

Roger S. Day, Department of Biomedical Informatics, University of Pittsburgh

Despite an explosion of research to exploit biomarkers for clinical application in diagnosis, prediction and prognosis, the impact of biomarkers on clinical practice has been limited. The elusiveness of clinical utility may partly originate when validation studies are planned, from a failure to articulate precisely how the biomarker, if successful, will improve clinical decision-making for patients. Clarifying what performance would suffice if the test is to improve medical care makes it possible to design meaningful validation studies. But methods for tackling this part of validation study design are undeveloped, because it demands uncomfortable judgments about the relative values of good and bad outcomes resulting from a medical decision.
An unconventional use of "number needed to treat" (*NNT*) can structure communication for the trial design team, to elicit purely value-based outcome tradeoffs, conveyed as the endpoints of an *NNT* "discomfort range". The study biostatistician can convert the endpoints into desired predictive values, providing the criteria for designing a prospective validation study.

Next, a novel "contra-Bayes" theorem converts those predictive values into target sensitivity and specificity criteria, providing the basis to design a retrospective validation study.

In practice, *NNT*-guided dialogues have contributed to validation study planning by tying it closely to specific patient-oriented translational goals. The ultimate payoff comes after completing and reporting a well-justified study. Readers will understand better what the biomarker test has to offer patients, because the study provides a biomarker test decision framework directly aligned with the targeted clinical decision challenge.

Acknowledgements: These ideas developed during the course of study design discussions with clinical translational investigators G. Larry Maxwell, William Bigbee, Ali Zaidi and Larisa Geskin, whose patience and insights were instrumental. Important suggestions and discussions are due to the generous efforts of Richard Simon, Gregory Cooper, Yan Lin, Daniel Normolle, Nick Lange, and Paul Marantz. I gratefully acknowledge the support of NIH grants R01 LM 010144, P30 CA047904, P50 CA121973, and DOD grants W81XWH-11-2-0131 and W81XWH-05- 2-0005.

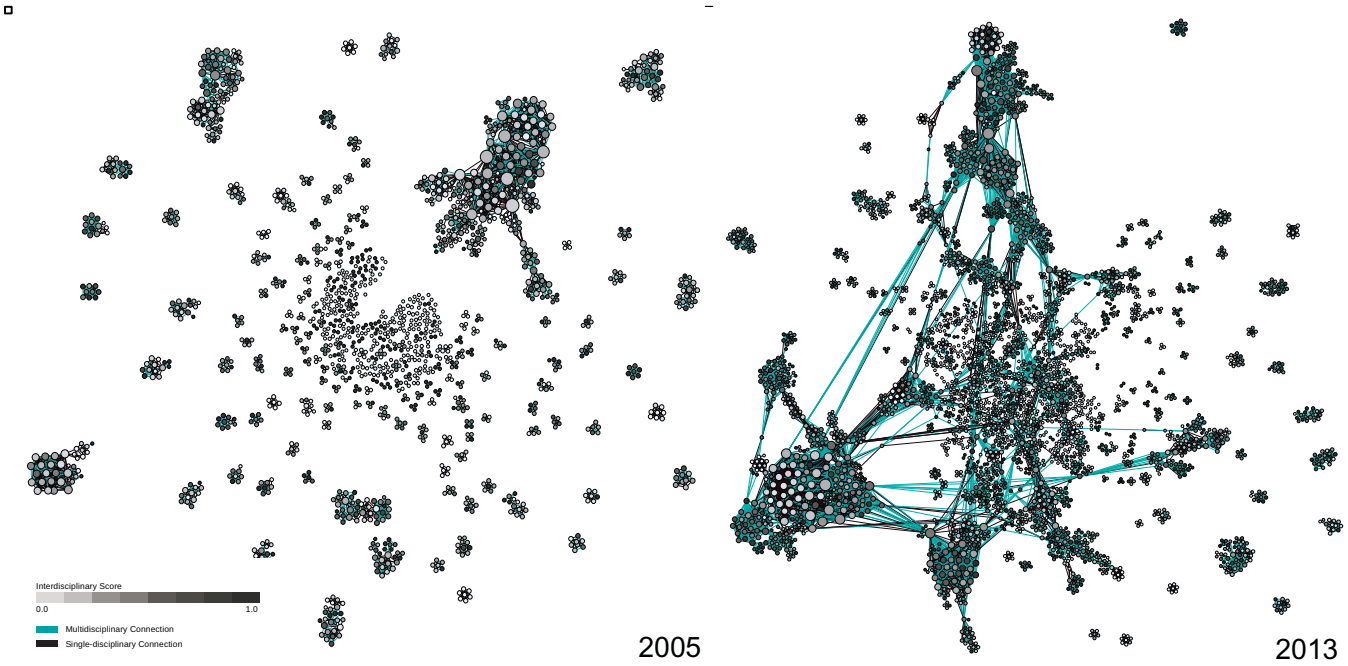
Spike in interdisciplinary research helps decode autism's complexity

Marlena Duda, BS; Maude M. David, PhD; Jae-Yoon Jung, PhD and Dennis P. Wall, PhD

Department of Pediatrics and Psychiatry (by courtesy), Division of Systems Medicine, Stanford University, Stanford, CA 94305

Summary: By mining the entirety of published autism studies we have created the Autism Research Connectome, an interactive online tool for the visualization of key research connections. This tool will allow investigators to discover potential interdisciplinary collaborators and in turn facilitate discovery towards the personalization of autism diagnosis and treatment.

Background: Autism is heterogeneous, complex, and arguably a condition of many conditions. Numerous questions regarding the etiology of ASDs remain unanswered, which can partly be attributed to compartmentalized research efforts that cannot fully address the multimodal causality of these disorders. **Methods:** We constructed the autism bibliome – a complete database of autism related research publications from the last 15 years. In turn, we created the Autism Research Connectome to visualize research connections between individuals, specifically to show the progression of this research network over time. Through text mining and natural language processing of bibliome keywords and abstracts, we classified research collaborations between autism investigators as either interdisciplinary or single-disciplinary. **Results:** There has not only been a steady increase in research efforts geared towards autism over the last 15 years, but also a shift towards more collaborative studies. Interdisciplinary collaborations have increased more than eightfold since the year 2000, and over 84% of authors active in 2014 are publishing interdisciplinary work. Furthermore, we found interdisciplinary collaborations to be significantly more impactful than single disciplinary collaborations ($p << 0.05$). **Discussion:** The complex and heterogeneous nature of autism has necessitated a transition away from traditional methods towards a more cooperative and interdisciplinary research landscape. Our results suggest that most, if not all, areas of autism research are starting to converge, and these convergences are leading not only to a richer research network but also to a causal network for autism. This network can, and likely will, lead to a way to decode autism spectrum disorder into its various subcomponents, and enable improved and increasingly more personalized approaches for both the detection and treatment of the many different forms of autism.



Immune Modulators in Diagnostics and Therapeutics: an Integrated Knowledge Approach

Nophar Geifman (PhD), Sanchita Bhattacharya (PhD) and Atul Butte (MD, PhD)
Dep. of Pediatrics, Division of Systems Medicine, Stanford University, Stanford CA

We present the development of a novel knowledgebase of integrated immune-related information and its extensive population with text-mined data, curated knowledge and experimental data. The usefulness of the knowledgebase in gaining new insights into the role immune modulators have in disease is demonstrated through several examples and use cases.

ABSTRACT

Cytokines and immune cells play a central role in both health and disease. Cytokines modulate immune responses and have been implicated both as diagnostics markers and therapeutics targets, while different cell types, such as white blood cells, have established immunological roles. To date, a systems level approach for integration and examination of immune patterns, such as cytokine level measurements along with blood cell counts, has not been applied in the context of disease.

Here we present a cytokine-centered knowledgebase focusing on the role cytokines play in disease, diagnostics and therapeutics. In this knowledgebase, cytokines are linked to a variety of different diseases and immune cell types. As part of its initial development, the knowledgebase was extensively populated with information mined from over 2.4 million PubMed abstracts along with curated textbook knowledge and data from gene expression arrays. As an innovation in the approach to knowledgebase development, this knowledge model is designed to incorporate cytokine and other immune-related patterns mined from open clinical trials data, electronic health records, genome wide association studies and other large-scale experimental data.

The usefulness of the knowledgebase for obtaining new insight regarding cytokines, cell-types and disease is demonstrated through several examples. Clustering of cytokine-disease co-occurrences from biomedical literature is shown to capture current medical views, as well as potentially interesting relationships between diseases. The ability of the text-derived data to capture immunological trends is further demonstrated by a high degree of overlap between this dataset and a network generated from curated textbook-derived data. Correlating cytokine-gene expression in a variety of diseases reveals diseases with similar cytokine expression patterns which could potentially share similar underling mechanisms. Finally, we demonstrate how the knowledgebase can be used to examine cytokines and cells in their use as diagnostic measures and as therapeutic agents.

To conclude, this knowledgebase of integrated Big Data along with curated knowledge holds potential for new discoveries in diagnostics and therapeutics. Using integrated large-scale data such as data mined from scientific literature, experimental and clinical data, a better understanding of the immune mechanisms underlying disease can be achieved and applied to diagnostics and therapeutics.

Identification of Causal Cascades and Unobserved Intermediate Structure in Input-Output Systems.

Clark Glymour, Ph.D. Alexander Murray-Watters , M.S.
Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, PA 15213, USA

Summary We describe a fast algorithm, provably asymptotically correct under specified assumptions, for identifying causal structure in systems with inputs, outputs and unmeasured, one-layer intermediate variables with or without feedback cycles. An example is given for modeling cellular signaling pathways from genomic alterations (observed) that lead to aberrations in signaling pathways (latent), and the latter in turn lead to alterations to gene expression transcripts (observed). The model was obtained from a cancer genomic dataset using 4000+ observed variables.

Background Machine learning for causal inference has developed fast, correct procedures for a variety of situations, including methods for identifying exogenous unobserved common causes and (some of) their causal relations, but no published method will identify unobserved intermediate variables and their causal relations—e.g., unmeasured cellular signal transduction systems that are produced by a causal cascade involving genomic alterations, aberration in signal transduction systems, and gene expression.

Methods We combine non-parametric strategies exploiting d-separation in graphical causal models and, to identify cycles, combine them with partially linear methods exploiting rank constraints on correlation matrices (Murray-Watters and Glymour, forthcoming; Kummerfeld, et al. forthcoming).

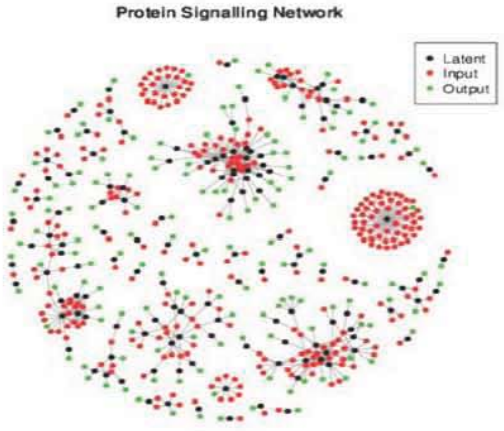
Results For a variety of simulated input-output systems, the procedure performs more accurately than factor analysis, the most commonly used procedure for identifying latent common causes. We also applied the procedure—disallowing cycles—to 4, 369 variables selected from more than 30,000 variables by highly significant LASSO regressions of expressions on gene alterations in a cancer dataset (http://cancergenome.nih.gov/) consisting of 562 observations (patients), 17,610 genomic alteration variables (recording whether or not a gene was altered), and 12,042 gene expression variables (originally continuous measures of mRNA levels, which were then converted into ordinal variables).

Discussion The results were produced on a single core laptop in 43 minutes; considerable speed-ups are possible with multiple cores and faster processors, making feasible the analysis of much larger variable sets. Current work involves applying an extended method that allows cycles to subsets of the data for which there are independently validated pathways.

References

Murray-Watters, A. and Glymour, C. (forthcoming). Discovering Nature's Hidden Springs, *Philosophy of Science*.

Kummerfeld, E., Ramsey, J., Yang, R., Spirtes, P., and Scheines, R. (forthcoming). Causal Clustering for 2-Factor Measurement Models. *Proceedings of the 2014 European Conference on Machine Learning*.



Acknowledgements
This research is undertaken under the auspices of the - University of Pittsburgh-Carnegie Mellon Center for Causal Discovery, supported by the National Institutes of Health under Award Number U54HG008540. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additional support was received from the James. S. McDonnell Foundation. Thanks also to Dr. Xinghua Lu, and Dr. Gregory Cooper, both of the University of Pittsburgh

Ranking adverse drug reactions with crowdsourcing

Assaf Gottlieb[1], Robert Hoehndorf[2], Michel Dumontier[3] & Russ B. Altman[1,4]

[1] Department of Genetics, Stanford University; [2] Department of Computer Science, University of Aberystwyth; [3] Stanford Center for Biomedical Informatics Research, Stanford University; [4] Department of Bioengineering, Stanford University

There is no publicly available resource that provides the relative severity of adverse drug reactions (ADRs). Such a resource would be useful for several applications, including assessment of the risks and benefits of individual drugs or drug classes and for personalized treatment. It could also be used to triage predictions of drug adverse events. We have used internet-based crowdsourcing to rank ADRs according to severity. We assigned 126,512 pairwise comparisons of ADRs to 2,589 Amazon Mechanical Turk workers and used these comparisons to rank order 2,929 ADRs. As expected, there is good correlation (ρ=0.53) between the mortality rates associated with ADRs and their rank. Our ranking underscores drug classes with excess types of severe ADRs and highlights severe drug-ADR predictions, such as cardiovascular ADRs for raloxifene and celecoxib. It also triages genes associated with severe ADRs such as epidermal growth-factor receptor (EGFR), associated with Glioblastoma multiforme and SCN1A, associated with Epilepsy. Finally, ADR ranking lays a first stepping stone in personalized drug risk assessment.

Pharmacovigilance plays a crucial role in the continuing evaluation of drug safety. Rofecoxib (Vioxx) (1, 2) and thalidomide (3) were withdrawn from the market because of severe adverse reactions (ADRs). ADRs contribute to excess length of hospitalization time, extra medical costs, and attributable mortality (4-6). Thus, assessment of the impact of ADRs on drug risk-benefit assessment has gained significant interest in recent years as several risk-benefit methodologies have been suggested for assessing drug safety and efficacy (7, 8). Two factors are essential for risk assessment: the prevalence of the ADR in the population (i.e. frequency) and the severity of the ADR in terms of medical (morbidity and mortality) or financial consequences. Risk estimates focus mainly on ADR frequency, as there is no publicly available resource that provides estimates of relative severity of ADRs. Thus, these methods either handle a single ADR at a time (7) or assign equal weights for all the drug ADRs (9). However, not all ADRs are of equal interest: life-threatening ADRs require more attention, while minor ADRs may not. Although a few severe life-threatening ADRs are well recognized, including liver failure, cardiac arrest, and others, there is presumably a gradation of severity from these down to the most benign. Of course, patients' subjective perception of the severity of an ADR varies widely, and so a ranking of ADRs is fundamentally a personal activity when it comes to individual patient decisions. Nonetheless, a ranking of ADRs based on perceived severity is a useful starting point for riskbenefit assessment or for personalization and is the focus of this paper.

Connecting the signatures: Associations between small molecules and microRNA in Malignant Melanoma

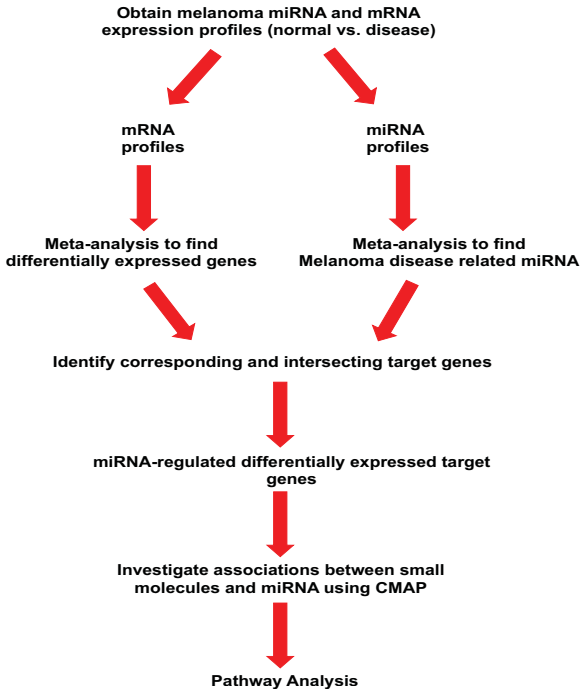
Authors: Anuvrat Jha[1], Kelly Regan[1], Phillip R.O. Payne, PhD[1]
Institutions: Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio[1]

Summary: The incidence of melanoma is rising faster than any other cancer in the United States. Current therapies for melanoma are limited in their efficacy and durability. MicroRNAs (miRNAs) have been implicated in different cancers and represent potential anti-cancer agents. In this study, we utilize the Connectivity Map to elucidate pathways connecting small molecules with differentially expressed miRNAs in melanoma.

Background: Melanoma is the most deadly form of skin cancer. In 2012, there were over 76,000 new cases of melanoma in the United Sates and nearly 10,000 deaths from this disease. Current therapies for melanoma are limited in their efficacy and durability. Moreover, creating novel drugs to combat melanoma is lengthy and costly. A drug takes 10 to 20 years from development to delivery to patients and costs, on average $1.2 billion. Furthermore, 90% of new drugs fail during the initial phase of development. Using drug repurposing, we hope to begin to address these issues. The Connectivity Map (CMAP), a drug repurposing database, identifies novel associations between diseases and drugs. These associations can be used to generate hypotheses which can be validated through *in vitro* and *in vivo* research. CMAP was created at the Broad Institute in order to link gene patterns associated with diseases with corresponding patterns produced by drug candidates[1]. CMAP was generated from treating MC7, PC3, HL60, and SKMEL5 (melanoma) cell lines with 1309 district drugs and DMSO (vehicle control for 6 hours on Affymetrix U133a microarrays[1]. Moreover, we will incorporate the LINCS database, which is similar to CMAP, and has data on chemical reagents (4400 drugs and bioactives) tested on melanoma cell line A375. The applications of CMAP were further enhanced by the association of drug-induced gene expression profiles to those miRNAs in a broad spectrum of human cancers and Alzheimer's disease[1,2].

Purpose: The aim for this project is to investigate the effects of miRNAs on malignant melanoma gene expression and elucidate pathways connecting these gene expression patterns with those of small molecules in order to discover novel therapies.

Methods:



Results: We have identified four mRNA expression profiles and three miRNA expression profiles in malignant melanoma from the Gene Expression Omnibus. We have performed meta-analysis on miRNA expression profiles, and identified 22 differently expressed miRNA among normal vs. melanoma patients (FDR <0.05). We are in the process of completing meta-analysis on the mRNA expression profiles.

Discussion: The significance of this study is to discover novel uses for existing small molecules. We expand the use of the systematic drug repurposing method, and are the first to apply it to cutaneous primary and metastatic melanoma. Future directions include incorporating datasets from TCGA and target gene identification. Small molecule associations through CMAP and LINCS and pathway analysis will follow. Future directions of this overall study will include validating findings from this *in silico* model in *in vitro* and *in vivo* experiments.

Web Links:
GEO - http://www.ncbi.nlm.nih.gov/geo/
TCGA- http://cancergenome.nih.gov/
LINCS- http://www.lincsproject.org/

References:

1. Meng, F., Dai, E., Yu, X., Zhang, Y., Chen, X., Liu, X., … Jiang, W. (2014). Constructing and characterizing a bioactive small molecule and microRNA association network for Alzheimer ' s disease. *Journal of Royal Society Interface*. doi: 10.1098/rsif.2013.1057.
2. Jiang, W., Chen, X., Liao, M., Li, W., Lian, B., Wang, L., … Li, X. (2012). Identification of links between small molecules and miRNAs in human cancers based on transcriptional responses. *Scientific Reports*, *2*, 282. doi:10.1038/srep00282

Systematic integrative analysis of immune pharmacology

Brian A. Kidd[1,2,5] PhD, Aleksandra Wroblewska[1,5] PhD, Mary R. Boland[4] MA, Judith Agudo[1] PhD, Miriam Merad[3] MD PhD, Nicholas P. Tatonetti[4] PhD, Brian D. Brown[1] PhD, and Joel T. Dudley[1,2] PhD

[1]Department of Genetics and Genomic Sciences,
[2]Icahn Institute for Genomics and Multiscale Biology,
[3]Department of Oncological Sciences,
Icahn School of Medicine at Mount Sinai, New York, NY, USA
[4]Department of Biomedical Informatics,
Columbia University Medical Center, New York, NY USA
[5]These authors contributed equally to the work
Corresponding authors: joel.dudley@mssm.edu or brian.brown@mssm.edu

Summary: Pharmaceutical drugs of all types and classes are associated with immune system perturbations but the mechanisms of these associations are often not understood. We developed an integrative computational analysis that provides a framework for understanding immune pharmacology and developing targeted immune cell therapies through systems pharmacology.

Introduction and Background: Global patterns of drug influences on the immune system are not well understood. Given the central importance of the immune system to human health and disease, this incomplete knowledge has serious consequence for treating disease and minimizing unwanted side effects. Here we developed and applied a novel integrative computational approach to perform a systematic analysis of potential interactions between >1,300 drugs and >250 immune cells. From these interactions we constructed an immunepharmacology map (IP-map) of >69,000 predicted drug-cell connections. These connections can enable improved understanding of immune pharmacology and provide a framework for developing novel approaches for modulating immune activity through systems pharmacology.

Methods: We constructed a matrix of predicted interactions between all pairwise combinations of 1,309 drugs and 304 immunological state changes using a rank-based, pattern-matching strategy that evaluates the overlap between the top and bottom ranked genes. The degree of overlapping genes quantifies the likely effect of a drug on an immune cell—the “immunemod score”. For each score, we obtained a measure of statistical significance by comparing the immunemod scores from the actual gene overlaps to a set of scores derived from a large number of permutations of the gene ranks. These permutations provide an estimate of the expected values for immunemod scores based on random overlaps and give a p-value for each immunemod score.

Results and Discussion: Overall, the IP-map identified 69,995 significant connections between drugs and immune cell states, which revealed many known and novel interactions. We demonstrated the utility of IP-map by showing we can correctly predict the manner of influence and specificity of the drugs clioquinol and guanfacine on neutrophil migration and regulatory T cell frequencies respectively. The IP-map also guided discovery of drug influence on immune cell frequencies in patient data from electronic medical records. Our method provides a new tool for screening thousands of interactions to identify novel relationships and provides guidelines for more rational manipulation of cells in the immune system.

IMPortal - Web portal to facilitate clinical decision support

Ravikumar K.E., Majid Rastegar-Mojarad, Saeed Mehrabi, Wagholikar K.B., and Liu, H
Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55901

Introduction: Since the inception of the Human Genome Project [REF] in 1990, genomic information has become an increasingly integral component in the diagnosis, treatment, management, and prevention of numerous diseases and conditions. While a large amount of IM knowledge information is available in curated databases such as OMIM or PharmGKB, much of this information remains 'locked' in the unstructured text of biomedical publications, particularly as new findings are published and certain findings are not incorporated into curated resources such as PharmGKB. Text mining approaches can accelerate the process of assembling IM knowledge in published literature. However, developing text mining systems with semantic understanding capability in the biomedical and clinical domains is very challenging. Integrating the scientific literature with other genomics databases resources and clinical EHR is an added dimension to this complexity.

Objective: The chief objective of IMPortal, a web portal for individualized medicine called is to support the decision making process in individualized medicine (IM) practice enabled by the use of advanced text mining techniques including information retrieval.

Informatics infrastructure of IMPortal

Figure 1 gives an overview of the informatics infrastructure of IMPortal. The architecture consists of following features: 1) Entity profiles consist of i) entity relations curated/extracted ii) evidence excerpts from original articles, and iii) relevant articles. 2) IM entities and entity relations are mined: i) existing records in curated knowledge bases and ii) entities mined from literature using information extraction systems (e.g. named entity recognition or event detection systems). 3) Excerpts are from articles containing the information related to the entities or relations. 4) Entity profiles are assembled using text mining techniques dynamically updated and serve as the knowledge repository for facilitating IM decision-making.

Figure 1 – Informatics infrastructure of IMPortal

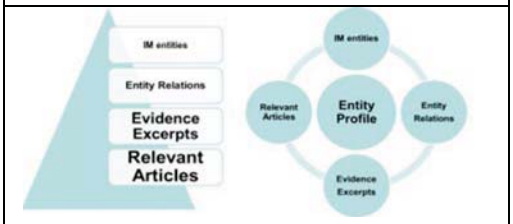
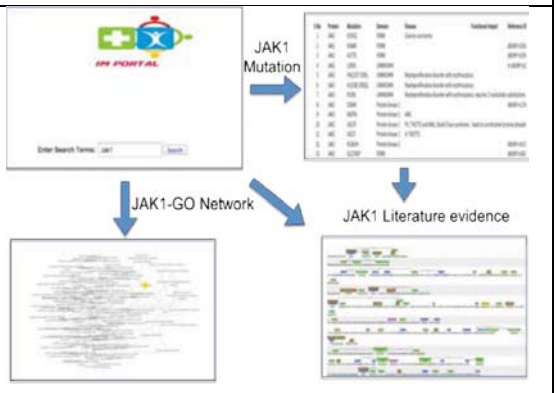


Figure 2 – IMPortal web architecture



Functionality of IMPortal: IMPortal has the following functionalities 1) Search through keywords and 2) Browsing through the Individualized Medicine (IM) entities and events 3) Cross-linking of information across multiple sources of genomics data. Figure 2 captures the overview of IMPortal web workflow

NLP in IMPortal: One unique feature of our NLP engine is its unique ability to enable semantic interoperability between text mining results with information stored in knowledge bases. IMPortal consists of a multi-layered NLP system, which represents information at different layers (documents, excerpts, sentences, entities and relations, and normalized entities and relations) so that downstream applications can interactively take advantage of the information extracted at different semantic levels, providing a robust way to bridge the semantic gap between text semantics and expert semantics. IMPortal is currently under alpha testing before roll out for clinical care at Mayo clinic.

A Metadata Framework for CLL Tissue Management Systems

A recommendation for the interoperability and reuse of cancer tissue core data

THE OHIO STATE UNIVERSITY
COLLEGE OF MEDICINE

Cartik Kothari, PhD; Tasneem Motiwala, PhD; Andrea Peabody, MS; Marjorie M. Kelley, RN, MS; Philip R.O. Payne, PhD
Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio

ABSTRACT

Research into the genetic origins of cancer depends upon the availability of tissue for assays and experiments. Tissue samples are crucial to all biomedical experiments. However, tissue repositories do not track the usage of tissue samples that are procured from them; there is no data linkage between experimental methods and the tissue samples that they use. We propose the implementation of metadata frameworks around tissue repositories that will link tissue samples with experiments as well as with the results of those experiments. This will enable the efficient usage of tissue samples that are typically expensive to procure and store, as well as facilitate the development of tissue centric data integration and analytical methodologies.

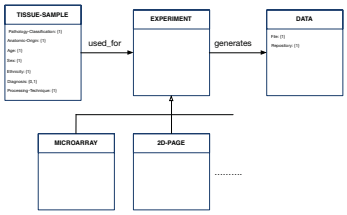
Background

- Approx. 16000 patients were diagnosed with Chronic Lymphocytic Leukemia (CLL) in the US in 2013 with ~4500 mortalities.
- Prognosis of patients varies from weeks to several years
- Underlying genetic causes are not well understood
- A research team at OSU pioneered one of the first methods to culture CLL tissue[1].
- The CLL Research Consortium (CRC), comprising 8 cancer research centers, uses a tissue management system as part of its information management system.
- At present, de-identified tissue data is stored with patient age, sex, and race, and a pathology report, where available.

- **Pathology:** Benign, Malignant, or Normal
- **Anatomic-Origin:**
- **Processing-Technique:** Frozen, Fresh, Paraffin-Embedded
- **Diagnosis:** Optional

Proposed Metadata Extensions

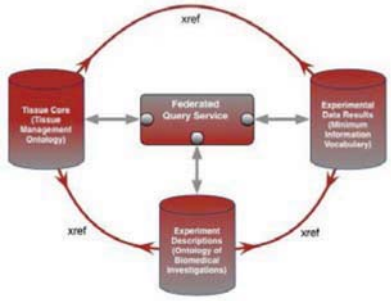
1. Cross reference tissue data with data obtained from experiments performed on the tissue samples, as well as the experiment details
   - For example, cross reference sample data with microarray data submitted to GEO from microarray experiments on sample
   - This is similar to open access journals requiring authors to submit their experiment data to open access knowledge bases
2. Capture geo-location from where tissue sample is sourced
3. Use metadata tags defined in W3C standard and other ontologies and vocabularies for
   - Tissue management
   - Experiments (Ontology for Biomedical Investigations)
   - Anatomical Structures (FMA)
   - NCI Metathesaurus/Diseasomes



Cross referencing tissue data with experiments and generated data facilitates:

1. Avoidance of redundant research
2. Comparison of experimental techniques and data analyses
3. Combination of results from various experimental techniques to obtain a "big picture"
4. Multivariate Data Analyses and Knowledge Discovery

Proposed Architecture



An architecture for tissue core data services that is cross referenced with data services for experimental descriptions and experimental results. The relevant ontology for describing each data service is shown in brackets. A federated query service integrates data obtained from each service endpoint interface

Example Query

Find microarray data from GEO that was generated from tissues extracted from lymph of female CLL patients between 50 and 65, sourced from San Diego

Concerns

- Combining Sample Location, Diagnosis, Sex, Age, and Race may help in narrowing down group of patients and defeat the purpose of anonymization
- Entirely new data sharing plans will need to be developed
- Regulatory Compliance
  - Data Sharing Agreements
  - Human Subjects Protections
  - Privacy Concerns

References

[1] Heerema, N. A. et al. Stimulation of Chronic Lymphocytic Leukemia (CLL) Cells with CpG Oligodeoxynucleotide (ODN) Gives Consistent Karyotypic Results among Laboratories: a CLL Research Consortium (CRC) study. Cancer Genetics and Cytogenetics. Dec 2010; 203(2): 134 – 140.

QWRAP: A Comprehensive Microbiome Analysis Workflow

Ranjit Kumar, Ph.D[1]; Travis Ptacek, Ph.D[1], Casey D. Morrow, Ph.D[2]; Elliot J. Lefkowitz, Ph.D[13]

[1] Biomedical Informatics, CCTS, University of Alabama at Birmingham, Birmingham, AL, USA
[2] Cell, Developmental and Integrative Biology, University of Alabama at Birmingham, Birmingham, AL, USA
[3] Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL, USA

Background : Next generation sequencing techniques such as 16S microbiome sequencing provide ways to explore the complex microbial communities in different sites of the body. Several human diseases have now been associated with changes in the human microbiome. Employing microbiome analysis in the clinical setting to support personalized approaches to treatment is on the horizon. Although several software tools are now available which perform microbiome data analysis, a major limitation is that their use requires significant computational knowledge. Sharing these data with basic researchers and clinical practitioners also requires significant education on their part to be able to interpret the results in a straightforward, meaningful manner. Therefore, in an effort to support analysis of microbiome data, and promote sharing and use of that data in a high-throughput environment supporting multiple projects and investigators, we developed a workflow called QWRAP.

Methods : QWRAP acts as a software wrapper, making use of several existing bioinformatics tools including QIIME, FASTQC, FASTX, USEARCH, and R. QWRAP scripts are written using a combination of bash, perl, python, HTML and R languages. QWRAP code is freely available at https://github.com/QWRAP/QWRAPv2. The QWRAP manual "QWRAP-Reamde.docx" provides instructions to process the single-end and paired-end datasets.

Results : QWRAP makes it easier to analyze multiple microbiome datasets simultaneously, supporting the needs of a local core facility that is involved in the analysis of multiple large-scale datasets. QWRAP runs in a LINUX environment and will efficiently utilize multiple nodes within a high performance-computing cluster. Following analysis, QWRAP generates a static HTML report, which can be used to inspect and share the results of the analysis. The report also includes information which guides the researcher or clinician in interpreting the results.

Discussion : Clinical Implications – QWRAP is currently used in our microbiome core facility and is analyzing hundreds of microbiome samples per week for numerous researchers and clinical investigators. Using QWRAP, we are able to achieve the analysis turnaround time of less than 24 hrs after receiving the raw data. Currently we are using QWRAP to analyze microbiome associations with various human diseases like cancers, H. *pylori* infection, C. *difficile* infection and metabolic disorders such as obesity and diabetes. To demonstrate the clinical utility of QWRAP, we analyzed the microbiome of patients with chronic *C. difficle* infection who were subsequently treated by a fecal transplant. Fecal samples were collected before, and at several times post-transplant for microbiome analysis. The QWRAP analytical workflow enabled basic and clinical investigators to analyze and compare the temporal changes in taxa and diversity of the patient's microbiomes following transplant.

The Impact of Reference for Proportion Prediction of Tumor Samples using DNA Methylation Deconvolution

Haiqing Li PhD
City of Hope, Duarte, CA

Abstract

Tumor sample represents a composition of different type of tumor specimens, such as tumor cell, immune cell, and fibroblasts. The heterogeneity of tumor sample could be a predication for cancer diagnostic. Computational deconvolution can be used to estimate the cell specific proportion from heterogeneous sample data. A well-established reference dataset, which represent the purified profile of each subset cell, is necessary to generate accurate estimation for most deconvolution analysis algorithms. This study presents the impact of reference on the prediction analysis of breast cancer using methylation data. A reference for breast cancer was proposed using public available data and was tested on patient samples from TCGA.

Background

The proportion of major cell types of a tumor sample is important information for cancer diagnostic. However, most current pathology techniques can only show the proportion from a small piece of the whole tumor. These diagnostic methods have restricted requirements for samples preparation, and they are costly. DNA Methylation is a relative reliable epigenetic marker and could be easily extracted from most type tumor samples. Using computational deconvolution analysis to estimate the cell type proportion from methylation data provides an overview of the whole tumor sample directly. Most deconvoluation algorithms request a reference data set, which represents the purified subtype cell profile(1).

Experiment and Result

This study establishes a reference for breast cancer deconvolution analysis using a set of methylation data of immune cell profile (GSE35069) and breast cancer cell line profile (GSE42944). The cell lines that cannot be clustered within each sub cell type were removed. These methylation data was processed using linear mix effect model(2) to extract the different methylated region (DMR) as reference for deconvolution analysis. The top 500 DMR was used as reference for deconvoluation analysis using quadratic programming method(2). 10 paired tumor/normal breast cancer patient samples with methylation data were collected from TCGA to test the reference. Figure 1 shows the prediction results. Tumor sample and normal sample shows different sub cell type proportion. Tumor samples have high proportion of Luminal subtype, which is not observed in normal type. However, the Basal A subtype are shows in both samples. This could be caused by the similarity of the signature between Basal A and normal tissue. CD8T proportion is also not significant due to the relative small size of the reference (only 2 CD8+ cell lines out of 43 immune cell lines in the reference).

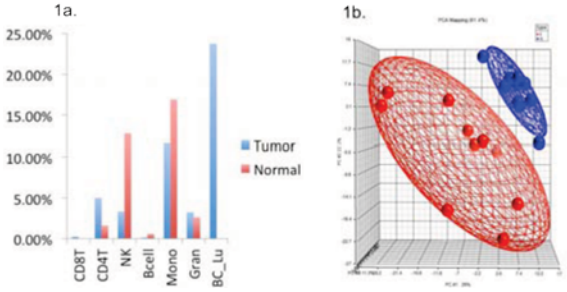


Figure 1. 1a) Estimation of cell proportion (mean) of 10 paired normal/tumor Breast Cancer Samples from TCGA; 1b) Cluster of methylation of breast cancer samples using top 500 DMR

Discussion

The methylation deconvolution analysis can predict the cell type proportion of the whole tumor sample. The reference data could impact the estimation result. For example, tumor tissue also contains normal cells. In order to observe the contamination from normal cells within tumor samples, normal samples should add to the reference list. Meanwhile, even the purified sub cell type data are available now, researchers need more detail meta data for normalization. A central data repository for this type reference data could benefit the cancer research using computational deconvoluation method.

Reference

1. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013;25(5):571-8. doi: 10.1016/j.coi.2013.09.015. PubMed PMID: 24148234; PMCID: PMC3874291.

2. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86. doi: 10.1186/1471-2105-13-86. PubMed PMID: 22568884; PMCID: PMC3532182.

Comparing 90-Day and 180-Day Hospitalization Risk Models

Anuj Misra, BA, Andrew M. Heekin, Ph.D., Charis Kaskiris, MS, Jakka Sairamesh, Ph.D.
Advisory Board Company, Austin, TX

Learning Objective

Understand the benefits in predicting hospitalizations, within a certain time period, based on historical patient information.

Abstract

Risk stratification models were deployed at large health system to predict hospitalization within a 90-day time period and 180-day time period of a previous encounter. Overall, the models performed well; they achieved an accuracy rate of 79% and 69%, respectively, for Tenant A and Tenant B. A majority of patients who were hospitalized within 180 days of their previous encounter returned within the first 90 of those 180 days.

Introduction and Background

For quality of care purposes, it would be advantageous to know whether a patient is likely to have an unplanned hospital visit after an initial encounter. We present results of statistical models, which are useful in predicting hospitalizations.

Methods

We chose two de-identified data sets (tenants) of encounter-level data consisting of adults, 18 and older; Tenant A had 1,167,837 encounters and Tenant B had 1,386,221 encounters. Logistic regression was used to predict 90-day and 180-day hospitalization, based on 29 factors derived from each encounter including: age, gender, primary ICD9 diagnosis code, and prior medical history factors. 70% of data was randomly selected to train models at Tenant A, while the remaining 30% was selected for testing model accuracy At Tenant B 100% of the data was used for validation.

Results

Several metrics were used to measure performance and are shown in Table 1. Both the 90-day and 180-day model had a minimum area under the curve of 70%.

Table 1. Model performance

	Sensitivity	**Specificity**	**Precision**	**Accuracy**	**AUC**
Tenant A 90-Day	84%	72%	80%	79%	87%
Tenant A 180-Day	83%	72%	82%	79%	86%
Tenant B 90-Day	58%	73%	49%	69%	71%
Tenant B 180-Day	57%	73%	53%	68%	70%

Discussion

About 90% of patients hospitalized within 180 days were also hospitalized within the first 90 of those 180 days. The factors contributing most to hospitalization included: age, number of inpatient claims in prior 30 days, primary ICD9 diagnosis code, and number of days since previous encounter.

Conclusion

Hospitalization is costly for patients and healthcare providers. Identifying those at high risk of hospitalization would be a key step in lowering costs and improving patient outcomes in the long term.

Iterative Interactive Enrichment (IIE) of Natural Language Processing (NLP) of Patient Electronic Medical Records (EMR) Shared Across Institutions

Joyce Niland PhD[1], Rebecca Ottesen MS[1], Isaac Kunz MS[2], Janet Nikowitz[1], Weizhong Zhu PhD[1], Leanne Goldstein DrPH[1], Tarik Courdy[2], Mike Chang PhD[1], Ajay Shah PhD[1], Samir Courdy[2]; [1] City of Hope, Duarte, CA; [2] Huntsman Cancer Institute, Salt Lake City, UT

Summary

Much of the useful EMR patient information is contained in unstructured text. The IIE NLP project goal is to evaluate extensibility, portability and accuracy of shared NLP queries across institutions. Linguamatics queries were developed to extract coded information from complex pathology dictations. Query enhancement iterations were assessed through incremental processes.

Introduction

Obtaining analyzable information from EMRs is a critical investment, as patient data are collected in both structured and unstructured forms. Manual data abstraction analysis is cumbersome and costly; NLP applied to patient records can automate data collection of certain fields, allowing abstractors increased time to collect more complex data. Linguamatics worked with the Huntsman Cancer Institute (HCI) to develop queries to identify immunohistochemistry (IHC) marker results (positive/negative) from unstructured pathology dictations on malignant hematology patients. Queries developed in Linguamatics' I2E platform were then exported for independent deployment at City of Hope (COH), to assess exportability and reusability of queries developed at one institution to extract similar data at another institution. We applied an Iterative Interactive Enrichment (IIE) process to improve the IHC queries across COH and HCI, documenting the impact on completeness and accuracy of information extraction, and the most critical NLP features that impact these results.

Methods

To quantify extensibility and portability of NLP queries across institutions, several measures were assessed, including time and effort for the iterations of the project, degree of customization needed to export queries to another institution, and accuracy. The project was divided into two stages, with several iterations of query development within each stage. The first stage used a training data set to refine the query design, then enhancing the ontology with UMLS concepts. The training data set was compared to a codified curated data source as the 'gold standard' to estimate accuracy. The second stage used a test data set to assess utility of queries on pathology reports beyond the training data set. Precision, recall, and F score were calculated to quantify improvement with each query enhancement. At each iteration, results were reviewed to develop suggestions for query refinement.

Results

At HCI initial precision and recall in 140 marker results were measured at 66.7% and 8.6% respectively. Initial queries were exported to COH to evaluate against the gold standard data source without any modification, run on 10 IHC markers in 169 Non Hodgkin Lymphoma (NHL) patients. Only 56.8% were recognized as having at least one marker result by the NLP query. Of 207 marker result pair findings in the gold standard data set, precision and recall were measured at 89.0% and 57.4% respectively. After the first iteration of query improvements, precision and recall were 56.8% and 39.1% at HCI, and 82.5% and 68.8% at COH. The impact of additional iterative improvements currently in progress will be reported.

Discussion

Sharing NLP queries across institutions depends on enriching NLP patterns and ontologies to index documents. Initial testing showed that while language expression, phrasing and structure of documents may differ, moderate enhancement of queries for center-specific use generally improved recall rates with a modest decrease in precision. There will always be a need for human adjudication of more complicated clinical dictations. Ontology-based query enhancements add benefit across centers in that synonyms can lead to improved precision and recall measurements. Given the predilection among MDs for continued use of dictation as a "rapid fire" way to express patient care results, using NLP to provide a real-time feedback loop of extracted results could greatly enhance the accuracy and utility of codified data from patient records. Further, we have shown that institutions can benefit from each other in qualitative process improvements of extracting concepts via NLP via the IIE process.

Motif-based network generation towards drug repositioning – an experimental study for Diabetes

Iyanuoluwa Odebode[1], Aryya Gangopadhyay, PhD[1], Qian Zhu, PhD[1]
[1]Department of Information Systems, University of Maryland Baltimore County, Baltimore, Maryland, USA

Introduction:

With the advance in computational technologies, computer aided drug repositioning is a more effective way to find new usages of "old" drugs. However, a majority of studies for drug repositioning are targeting on a large amount of diverse types of data that may introduce more false positives and increase computational cost. In this study, we introduced a motif-based network with dramatically decreasing size compared to the original network by extracting diabetes relevant biological associations from CIDeR[1] (Curated Information of Disease Related Interactions) to further support drug repositioning.

Methods:

Diabetes based network extraction We extracted diabetes relevant interactions along with semantic types from CIDeR containing manually curated information consisting of 18,000 interactions from literature by searching for a term "Diabetes". Subsequently, we generated a network by integrating the above interactions.

Network motif discovery Inspired by one published study[2] to reveal distinct association patterns by applying network-based analysis, we explored FANMOD[3] to perform network motif discovery by limiting motif with 3 nodes. More specifically, the diabetes-based network was translated to an semantic type based interaction matrix as an input file for FANMOD. The FANMOD generates motifs along with statistical results that can be applied to determine the significance of those motifs.

Network motif based seed network generation As the network motifs illustrate the essential interactions among different types of concepts, including drug, gene, disease, SNP, etc., the network consists of such interactions significant to diabetes will provides novel insights for drug repositioning. Thus we mapped the selected network motifs to original diabetes network and aggregated those mapped sub-network into a network motif-based network to support further drug repositioning.

Results:

The diabetes network contains 6,871 interactions, which includes 1,853 nodes and 5,377 edges. We run this network by using FANMOD to identify network motifs. Finally we identified 34 motifs with P-Value less than 0.05, including "gene-disease-cellular compound" with P-Value = 0.034 and "SNP-gene-drug" with P-Value = 0.015. Subsequently, we generated a motif-based network by aggregating those network motifs for further supporting drug repositioning.

Conclusions:

In this study, we introduced a network motif based network towards drug repositioning for diabetes by applying network motif discovery, which can dramatically decrease the size of the network and extract a sub-network significantly relevant to diabetes. Drug repositioning candidate can be identified from this motif-based network consequently. In the next step, we will apply network-based analysis, for instance, perturbation study to identify the high-influence nodes in the seed network and subsequently determine drug candidates.

Reference:

1. Lechner, Martin, et al. "CIDeR: multifactorial interaction networks in human diseases." *Genome Biol* 13 (2012): R62.
2. Zhang, Yuji, et al. "Network-based analysis reveals distinct association patterns in a semantic MEDLINE-based drug-disease-gene network." Journal of biomedical semantics 5.1 (2014): 33.
3. Wernicke, Sebastian, and Florian Rasche. "FANMOD: a tool for fast network motif detection." *Bioinformatics* 22.9 (2006): 1152-1153.

Integrated database and knowledge database for genomic prospective cohort study in Tohoku Medical Megabank

Soichi Ogishima, Takako Takai, Kazuro Shimokawa, Satoshi Nagaie,
Hiroshi Tanaka, Jun Nakaya

Summary

The Tohoku Medical Megabank project is a national project to restoration of the disaster area in the Tohoku region by the Great East Japan Earthquake, and have conducted large-scale prospective genome-cohort study. Along with prospective genome-cohort study, we have developed integrated database and knowledge database which will be key database for realizing personalized prevention.

Leading Objective

- Integrated database and knowledge database for genomic prospective cohort study in Tohoku Medical Megabank

Abstract

The Tohoku Medical Megabank project is a national project to restoration of the disaster area in the Tohoku region by the Great East Japan Earthquake. It aims to become a center for the restoration in medicine of the entire Tohoku region by developing electronic network of medical records and by conducting large-scale prospective genome-cohort study. Along with prospective genome-cohort study, we develop an exceptional biobank in the Tohoku region which contributes to the restoration of medical cares in the disaster area and revitalizes related industries.

In our prospective cohort study, we will recruit 150,000 people, and collect (1) sample (blood, urine) (2) baseline and follow-up data. As for follow-up, we will collect clinical data as electronic health records provided by Miyagi Medical and Welfare Information Network (MMWIN). Collected data including laboratory measured data (genomic and omics data) have been de-identified and stored in a database, and have been integrated in our integrated database. We have also developed knowledge database to store known or newly revealed knowledge in our project. Knowledge database is RDFs-based database; triple store stores knowledge among URIs in integrated database, that is, genetic, habit, environmental and clinical factors leading to pathogenesis. We believe our integrated database and knowledge database will be key database for realizing personalized prevention.

Multi-dimensional nosology of drugs to prioritize novel drug uses

Hyojung Paik[1], and Atul J. Butte[1, *]

[1]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305 and Lucile Packard Children's Hospital, Palo Alto CA 94304, USA

* Correspondence should be addressed to abutte@stanford.edu

Abstract

Drug repositioning refers to alternative drug use discoveries, which differ from the original indications of the drug. Suggesting novel drug uses using a guilt-by-association approach is well known method for drug repositioning. One challenge in these efforts lies in choosing which indication to prospectively test a drug of interest by capturing similarities of drugs. We systematically evaluated drug-drug relationships using phenotype-based view, such as therapeutic terms and side-effects, and genetic signatures from gene expression profiles for each drugs in order to address this challenge. Compared with control drug classifications, the our novel drug clusters reveals new use of drugs as well known therapeutic options for target disease.

Assessment of the Development and Delivery of Genomic Medicine Information Resources

Luke V. Rasmussen[1], Casey L. Overby[2,11], John Connolly[3], Christopher G. Chute[4], Joshua C. Denny[5], Robert R. Freimuth[4], Andrea L. Hartzler[6], RoseMary Hedberg[1], Ingrid Holm[7], Shannon Manzi[7], Jyotishman Pathak[4], Peggy L. Peissig[8], Brian Shirts[9], Maureen Smith[1], Elena Stoffel[10], Peter Tarczy-Hornoch[9], Marc S. Williams[11], Wendy A. Wolf[7], Justin B. Starren[1]

[1]Northwestern University Feinberg School of Medicine, Chicago, IL; [2]University of Maryland School of Medicine, Baltimore, MD; [3]The Children's Hospital of Philadelphia, Philadelphia, PA; [4]Mayo Clinic, Rochester, MN; [5]Vanderbilt University, Nashville, TN; [6]Group Health Research Institute, Seattle, WA; [7]Boston Children's Hospital, Boston, MA; [8]Marshfield Clinic Research Foundation, Marshfield, WI; [9]University of Washington, Seattle, WA; [10]University of Michigan, Ann Arbor, MI; [11]Geisinger Health System, Danville, PA;

Abstract: *Genomic medicine introduces new challenges to providers and patients, including communication of rapidly changing knowledge about the relationship between the human genome and health. One approach is the use of information resources, either proprietary or publicly available, however this approach requires a commitment to manage and disseminate this evolving content. We present the results of a survey of ten academic medical centers and health systems to assess strategies and opinions concerning the management of information resources for genomic medicine.*

Introduction: Advances in genomic medicine bring new challenges to both providers and patients, who are faced with not only a growing amount of new genetic data to process but also continual changes in the understanding of how this data should be interpreted. As a form of passive decision support, information resources (e.g., clinical references, patient education materials) can be offered to both patients and providers, but challenges exist for how they should be developed, distributed and maintained (e.g. using existing or developing new resources, ensuring content is reviewed and updated). Institutions considering implementation of genomic medicine programs must consider their strategy for providing information resources. To examine this challenge, we explored the current and future strategies at ten academic medical centers and health systems in varying stages of implementation of genomic medicine programs.

Methods: The electronic Medical Record and Genomics (eMERGE) Network and Clinical Sequencing Exploratory Research (CSER) consortium deployed a survey to capture medical center and health systems' ("sites'") current and future plans for providing information resources to patients and providers, modes of delivery (i.e. external website, embedded as a resource in the electronic health record (EHR), paper handout), opinions on the state of existing resources, and how generalizable resources are across institutions. Questions differentiated between provider and patient resources to assess differences between these audiences. Additionally, eMERGE sites were asked to provide a separate set of responses for pharmacogenomic (PGx) and other genomic medicine (GM) scenarios, if they anticipated a difference in the delivery of resources or opinion about available resources.

Results: In total, 12 responses were received from eight eMERGE sites (two sites reported separately for PGx and GM) and two CSER sites. Of the 12 responses, 10 indicated current or future plans to provide information resources. Mode of delivery for providers is primarily the EHR (9/10) and a content management system (CMS) or website (6/10). Mode of delivery for patients is primarily paper/pamphlet handouts (8/10), and a CMS or website (8/10). With respect to internally versus externally hosted resources (those managed by the site or an external party), most respondents indicated that they have plans to or already host content internally (10/12 for both provider and patient resources). Few sites plan to host content externally (3/12 for providers and 1/12 for patient resources), including hosting both internal and external content (3/12 for provider and 1/12 for patient resources). Respondents indicated strong or moderate agreement that existing resources support patients (5/12) and providers (6/12), however there was also agreement about the need to create new content (11/12 for both). Finally, half of responses indicated moderate to strong agreement that provider resources should be site-specific and the majority (11/12) agreed that resources are generalizable to other sites. Responses for patient resources were similar (6/12 and 10/12, respectively).

Discussion: Survey findings show differences in how information resources for genomic medicine are delivered to providers and patients at several academic medical centers and health systems in the U.S. In addition, we found a trend indicating that while portions of the content may be generalizable, there is a perceived need to have site-specific information available. These results identify current usage of and gaps in information resources that can inform further development.

Acknowledgements: eMERGE and CSER sites are supported by multiple research grants from the NHGRI.

Gene expression-based "connectivity mapping" algorithms identify novel synergistic, anti-resistance drug combinations for use in malignant melanoma

Authors: Kelly Regan[1], Alex Mysiw[1], Phillip R.O. Payne, PhD[1]
Institution: Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio[1]

Summary

We present the first application and evaluation of gene expression-based "connectivity mapping" enrichment statistics in a combined anti-drug resistance and mechanistic synergy model of drug combination predictions in malignant melanoma tumors.

Background

Bioinformatics-based drug repositioning involves finding novel indications for existing drugs, and includes computational, molecularly based "connectivity mapping" methods between gene expression signatures of disease states and drug-treated cell lines. The Connectivity Map is a reference database of Affymetrix U133A microarray experiments of four cancer cell lines (MCF7, PC3, HL60, SKMEL5) treated with 1,309 drugs and DMSO vehicle controls at a median concentration of 0.0001M for 6 hours uniformly. The basic assumption of the Connectivity Map is that aberrations in cellular gene expression reflect important mechanisms in cancer progression and responsiveness to drugs. Connectivity scores are calculated from a Kolmogorov-Smirnov (K-S) statistic. The desired output of these analyses is to identify negative connectivity scores, which imply an anti-correlated relationship between the drug and disease state. Other statistical methods for drug connectivity mapping, including Spearman's correlation, Fischer's Exact test, Wilcoxon rank-sum test, weighted Pearson correlation, logistic regression, eXtreme cosine (XCos) have been put forth; however a formal analysis of diverse enrichment statistics and combinations thereof, have yet to be studied.

Methods

A feasible approach for predicting drug combinations from connectivity mapping is to query gene signatures derived from drug resistant phenotypes in order to discover drug treatments that can potentially reverse resistance and sensitize cells to the desired drug. Additionally, drugs that can affect distinct gene expression modules or pathways of disease may also have synergistic effects when used in combination. These notions served as the basis for our integrative model of drug combination predictions using gene expression data from drug-treated cell lines from the Connectivity Map.

Results

We obtained publicly available microarray data from distinct studies of patient-derived melanoma tumors: 1) differentially expressed genes from dabrafenib-resistant vs. responsive melanoma tumors (both pre- and post-treatment) and 2) a meta-analysis of differentially expressed genes from three studies of metastatic melanoma vs. primary tumors. Results from using the Connectivity Map K-S statistic enrichment analysis at a permutation p-value ≤ 0.05 provided drug predictions to oppose metastasis and resistance to dabrafenib treatment in both the pre- and post-treatment melanoma tumors. Additionally, we have obtained results using other connectivity mapping methods, including sscmap and DrugVsDisease in order to generate meta-models combining these different scoring metrics. Further, we are in the process of employing functional module and pathway-based algorithms to corroborate traditional gene signature-based connectivity mapping methods.

Discussion

We report novel predictions of drugs to be paired with dabrafenib that have anti-resistance and anti-metastasis properties. In future work, we will investigate the use of *in silico* methods to generate drug combinations independent of a specific drug resistance model in melanoma. Future validation studies will be carried out with retrospective clinical informatics studies of electronic health records of patients treated with predicted drugs, as well as *in vitro* and *in vivo* experiments to test drug efficacy, including cell viability and cell migration assays.

Extracting a Concept Hierarchy from UMLS to Support Hierarchical Expansion in Database Search

Alexander J. W. Richardson, Emily A. Fireman, Hyeoneui Kim, RN, MPH, PhD,
Division of Biomedical Informatics, University of California San Diego, La Jolla, CA

Abstract

To improve search performance of PhenDisco, a new search platform for dbGaP, through hierarchical expansion, we developed algorithms to extract concept hierarchies from the UMLS Metathesaurus. Upon testing through creation of a small hierarchy using this algorithm, we found the majority of the erroneous relationships were caused by UMLS itself.

Introduction

PhenDisco is a new search platform developed to improve the search performance for dbGaP (the database of Genotypes and Phenotypes, http://www.ncbi.nlm.nih.gov/gap) through concept-based search where the synonyms and children concepts of the given search term are automatically included in search. Currently, PhenDisco only supports the former (i.e., synonym expansion). To implement the latter (i.e., hierarchical expansion) we first grouped the phenotype variables in dbGaP into 16 categories, which will serve as 16 "axes" of the PhenDisco concept hierarchy. Then we developed the algorithms that extract the concept hierarchy from UMLS, which are relevant to the concepts of the phenotype variables in dbGaP.

Materials and Methods

The hierarchy extraction algorithms were written in Python and consist of three sequential steps. First, *relationship harvesting,* involves gathering all hierarchical relationships relevant to the input concepts by querying the concept relations table of UMLS. In a prior work, we developed a pipeline that maps key concepts of the phenotype variables in dbGaP to UMLS Metathesurus (http://PhenDisco.ucsd.edu). Thus, the input concepts for the algorithms are the CUIs (Concept Unique Identifiers) mapped from the phenotype variables. The parent concepts of the input CUIs are retrieved and saved, and then the parents' parent concepts are retrieved and saved. This process continues until all hierarchical relations are gathered. Next, *node legitimacy testing* enforces the "two or more children policy" where all nodes in the hierarchy must have two or more sub-concepts. Finally, the *redundancy removal* algorithm eliminates the hierarchical relationships directly defined between children concepts and grant parent concepts. During the algorithm development, we observed looping relationships (i.e., child concept becomes a parent concept of its ancestor concept). Therefore, we wrote a function that catches the looping relationships. We also wrote a Python script that transforms the final hierarchy definition files into the Protégé-OWL format. We tested these algorithms using 100 disease concepts as input CUIs and two domain experts collaboratively reviewed the outputs.

Results

The 100 input CUIs were structured into a hierarchy using 190 subsumption relationships that satisfied the hierarchy extraction algorithms. Ten relationships were deemed problematic upon manual review. Nine relationships were caused by the erroneous hierarchy definitions in UMLS. One relationship was not incorrect but deemed too distant to each other by the reviewers. This was caused by deletion of the parent concepts in between due to the *node legitimacy testing* algorithm. Eleven input CUIs were not included in the hierarchy either because they do not have parent concepts or their parent concepts didn't pass the *node legitimacy testing*. Six looping relations were found.

Discussion and Conclusion

We developed hierarchy extraction algorithms that utilize existing hierarchies in UMLS to support concept-based search in PhenDisco. As the next step, we will apply our algorithms to generate concept hierarchies for the entire phenotype variables in dbGaP. Our algorithms caught many erroneous hierarchical relations in UMLS, which warrants a caution in adopting UMLS hierarchies for any concept hierarchy-based algorithmic data processing. This also indicates that our algorithms can be used to check the validity of concept hierarchies of interests.

Acknowledgement

This study was supported by NIH grant U54HL108460 and UH3HL108785.

Title

Sex-specific patterns and differences in dementia and Alzheimer's disease using informatics approaches

Authors

1) Jeremiah Geronimo Ronquillo, MD, MPH, MMSc; Geronimo Ronquillo LLC, Haymarket, Virginia
2) Merritt R. Baer, JD; Geronimo Ronquillo LLC; Haymarket, Virginia
3) William T. Lester, MD, MS; Laboratory of Computer Science, Massachusetts General Hospital, Boston, Massachusetts; Harvard Medical School, Boston, Massachusetts

Publication status: In press (manuscript accepted for 2015+ publication in Journal of Women and Aging)

Summary

Sex plays a critical but only partially explored role in the etiology, diagnosis, and prognosis of Alzheimer's disease. Applying informatics approaches to large neurodegenerative datasets can provide deeper insight into these differences. We report findings from one of the largest sex-stratified analyses of Alzheimer's disease and cognitive impairment to date.

Introduction and Background

Alzheimer's disease (AD) is a serious form of dementia that affects more than 5 million people in the United States, with the majority older than 65 years and nearly twice as many women affected as men. NIH has highlighted the critical need for explicitly addressing male/female differences in biomedical research, which would ultimately contribute to more personalized healthcare for the growing aging population. Clinical, medical, and genetic factors have partly explained the risk of developing cognitive impairment, but more effective prediction requires an integrated assessment of these types of modalities. To our knowledge, informatics approaches have not been applied to large aging and dementia research datasets to characterize important male/female differences.

Methods

The study population of patients originated from large clinical research datasets for cognitive impairment: the Alzheimer's Disease Neuroimaging Initiative, the National Alzheimer's Coordinating Center, and the Coalition Against Major Diseases. An informatics data pipeline in Python was created to extract existing fields, derive relevant factors, and standardize raw neurodegenerative data into a single population database. Statistical analyses included standard summary statistics and tests for significance. For the patient subpopulation with dementia, the probability of documented probable AD was modeled with logistic regression using candidate predictor variables representing important clinical, demographic, medical, or genetic factors.

Results

There were a total of 24270 patients with cognitive impairment, with 12737 (52.5%) females and 11533 (47.5%) males. Overall, 8138 (33.5%) patients were classified with mild cognitive impairment and 16132 (66.5%) with dementia, of which 12505 (77.5%) had documentation of probable AD. Analyses suggested females were 1.5 times more likely than males to have a documented diagnosis of probable AD, and several other factors fell along sex-specific lines and were possibly associated with severity of cognitive impairment.

Discussion

This study is one of the largest integrated modality analyses of sex-specific patterns and differences in cognitively impaired individuals to date. Our findings highlight the complex relationships between different sex-related factors and comorbidities in a large population of cognitively impaired individuals. As research efforts further intensify for age-related conditions such as dementia and Alzheimer's disease, we have a unique opportunity to capitalize on the increasing availability of data, sophisticated informatics tools, and public readiness to solve critical challenges facing healthcare today.

Rapid integration of cancer genomics data using Hadoop and Cloudera's Impala

Sittichoke Saisanit, Zayed Albertyn, Xing Yang, Padmanabha Udupa
Pharma Research and Early Development Informatics, Roche Innovation Center New York, 430 East 29th St, New York, NY 10016 USA

Abstract

Keywords: big data challenge, traditional solution, Impala, conclusion

The ever growing availability of next-generation sequencing (NGS) data from cancer samples and cell lines has enabled researchers to discover new cancer driver genes and biomarkers. Mutations, indels, gene fusions, along with sequence effect predictions have become increasing large. Each single source of data can be as big as multiple terabytes. Traditionally, we created a data warehouse using RDBMS to integrate these data for effective querying and analysis. We have explored an Apache Hadoop cluster using Cloudera platform as an alternative to relational database. We use Impala which is Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in Hadoop cluster. Impala takes advantage of Hadoop file system and offers querying capability to the data using standard SQL language. We found that creating an Impala table from Hadoop file system can be accomplished with relative ease using a single command. The speed of data loading is several magnitudes faster than SQL Loader. Once loaded, query performance including joins is also much faster than non-optimized relational database. In addition to direct query, we can also connect to Impala using Spotfire and Java API. We conclude that Cloudera Impala is a compelling alternative to RDBMS for quickly exploring genomics data. The speed and ease of data loading enables us to just load data without spending time on schema design, index creation, query tuning, data cleaning, and data transformation. Together with visualization capability of Spotfire, the platform allows data scientists to rapidly analyze data or simply performing various data queries to answer scientific questions in support of our drug discovery and development programs.

REACH NC Resource Finder: a multi-institutional portal of university resources to increase cross-disciplinary collaboration.

Sharlini Sankaran*, Ph.D., Hong Yi*, Ph.D., Marc Ciriello†, B.A., Michael Cherry†, B.Sc., Chris Barker‡, Ph.D., and Bhanu Bahl†, Ph.D.

*Renaissance Computing Institute (RENCI), and ‡School of Medicine, University of North Carolina Chapel Hill, Chapel Hill NC.
†Harvard Medical School, Harvard University, Cambridge MA

Summary (around 50 words): The REACH NC Resource Finder is a unique, multi-institutional portal of resources residing in North Carolina's academic institutions. This tool was developed as a collaboration of Harvard and UNC with a goal of implementing a customized portal to foster collaboration amongst researchers and with industry and economic development partners.

Background and introduction: Biomedical research today is highly collaborative and often multi-institutional; yet funding for resources and equipment remains relatively scarce. Oftentimes there is no central way to locate equipment and instrumentation that can be shared across departments, much less institutions. The REACH NC Resource Finder now provides a central, multi-institutional portal containing resources from across North Carolina academic institutions to make collaboration and equipment sharing easier. The Research, Engagement, And Capabilities Hub of North Carolina (REACH NC) is a publicly-accessible, searchable, web-based portal enabling quicker and easier location of statewide research expertise and assets. The Resource Finder is a multi-institution implementation of eagle-i, an open-source, open access application that makes it easy to discover biomedical resources at universities and research institutions. eagle-i is now entirely sustained under Harvard Medical School's CTSA grant. The project's goals have a strong national reach and value: Use of eagle-i has grown to over 35 Universities and is rapidly expanding. REACH NC and eagle-i have collaborated to launch the REACH NC Resource Finder based on a pilot project of the UNC Translational and Clinical Sciences (TraCS) Institute. The Resource Finder contains searchable listings of core facilities, instrumentation, and other assets across six public and private universities. The REACH NC Resource Finder is the first multi-institutional eagle-i network outside of the original central eagle-i network. As such, implementing the Resource Finder required extensive customization and set-up.

Methods: Implementing the Resource Finder involved two sets of decisions: policy and technical. Policy challenges required reaching consensus amongst REACH NC's institutional partners about what information should be included in such a portal and how the process could be standardized across institutions while ensuring that individual institutions retain control over their data. Technically, while setting up an institutional instance of eagle-i was straightforward, the question of how to set up a statewide network with multiple nodes had not been previously addressed. We decided to replicate the protocols and infrastructures implemented for the nationwide eagle-i search and adapt them towards a North Carolina-specific portal.

Results and Discussion: The REACH NC Resource Finder was launched in June 2014. It currently contains over 1,600 resources from six North Carolina institutions: Duke University, East Carolina University, North Carolina A&T State University, UNC Charlotte, UNC Greensboro, and UNC Chapel Hill. More institutions will be added to the Resource Finder in the coming months. The resources were also indexed to the national eagle-i network with the creation of a specialized "REACH NC Resource Finder" collection. A wiki page detailing the technical steps necessary to implement a similar portal is under development and will be made available to the open-source community. Next steps for the Resource Finder include integration with other REACH NC functionalities including the experts portal which uses the SciVal vendor software and currently contains over 10,000 researchers from 20 North Carolina institutions. We will explore automating or making more user-friendly the process of importing and exporting data. As more institutions join the Resource Finder, we hope to increase collaboration and resource-sharing. The resource finder is also useful for economic development by showcasing university assets: in one recent use case, the Resource Finder was used to identify core facilities and equipment related to a large economic development project that is currently under recruitment in North Carolina.

A Scalable Computing Platform for Human Genomics

Natasha Sefcovic[1], Andrew Clark[2], Christopher Gardner[2], Franklin Totten[3], Amandeep Chawla[3], Priyanka Oberoi[3], Sriram Sridhar[3], Ahsan Huda[1], Paul Hodor[3]
[1]Booz Allen Hamilton, Civil Commercial Group, Rockville, MD, [2]Booz Allen Hamilton, Civil Commercial Group, Boston, MA, [3]Booz Allen Hamilton, Strategic Innovation Group, Rockville & Linthicum, MD

Whole genomic sequencing of a large number of individuals is expected to have an unprecedented impact on health care. Potential benefits include identification of genetic factors that cause or are risk factors for specific diseases, discovery of novel therapeutic targets, and development of personalized treatment plans. As sequencing costs have been decreasing by orders of magnitude in recent years, it is becoming feasible that an individual's genome will be included in their medical record, setting a foundation for precision medicine. At the same time, this cost reduction is creating a boom in the number of available complete human genome sequences. The rate of sequencing data production has far outpaced the rate of growth in computing speed and digital storage capacity. Data analysis is turning into a bottleneck. New computational paradigms are needed for data analysis to keep pace with the abundance of experimental sequence data.

The abundance of data has created several specific challenges. The transfer of raw data may be too slow, such that it is impractical for data to be downloaded each time a new analysis is performed. Queries of large, traditional databases and retrieval of results can take too long. The algorithmic complexity of traditional analysis tools often does not scale well. A promising solution to address such challenges relies on distributed computing and cloud services. High performance, scalable systems can be built in a cost-effective fashion around a core of general-purpose computing tools and reusable components. They have the advantages of robustness, longevity, and portability.

Here we describe the development of a novel platform for the storage, management, and analysis of human genomic variants, built on open source tools from the Hadoop ecosystem. Hadoop software is widely used in a variety of applications involving massive datasets, and is known for its reliability and scalability. Amazon Web Services were used to deploy the system as a cloud service. As an initial use case, we focused on implementing a genome-wide association studies (GWAS) analysis pipeline. GWAS is of major importance in areas of human health such as pharmacogenomics and precision medicine. We developed data models for human genomic variation data in the NoSQL database, HBase. They capture single nucleotide polymorphism and indel information of genomic variants across the entire human genome, variant calls for each subject genome stored in the database, and pedigree information. The data models were optimized for fast retrieval of query results by massively parallel MapReduce operations. Queries support filtering by individuals and genomic region. Case and control cohorts are designated by the user. Genotype results from the queries are analyzed by PLINK wrapped in a MapReduce streaming job. We implemented a proof-of-concept, web-based user interface that includes submission of GWAS tasks and viewing of analysis results. Using data from the 1000 Genomes Project, we benchmarked the system's performance during the import, query, and analysis stages across datasets of increasing numbers of variants and individuals. Our findings

showed the feasibility of a Hadoop-based framework as an elastic, on-demand solution for performing large-scale GWAS analysis. The system can be expanded by integrating new and diverse dimensions of data from medical sources, and adding other use cases to enable a powerful, multi-purpose analysis platform. Development of such a platform will be critical to the deployment of genomic analysis in modern clinical applications, as well as for the development of novel research capabilities.

COmbined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL): A Robust Method for Selection of Cluster Number K

Timothy E Sweeney [1,2,†,*], Albert Chen [3,†], Olivier Gevaert [2]
1 – Department of Surgery, Stanford University
2 – Biomedical Informatics Research, Stanford University
3 – Stanford University
† - These authors contributed equally to this work
* - Corresponding author: tes17@stanford.edu

Abstract

Introduction

Multiple unsupervised clustering techniques exist, with multiple metrics by which the optimal number of clusters in a dataset (k) may be judged. Here we report a method, COmbined Mapping of Multiple clUsteriNg ALgorithms (COMMUNAL), that combines information from multiple clustering algorithms and multiple cluster metrics to enhance the robustness of a given assessment of k. We further show that the stability of a dataset's optimal k can be mapped by sequentially adding variables to the clustering algorithms.

Methods

Starting with datasets from three common cancers in the cancer genome atlas (TCGA) as well as generated Gaussian cluster datasets ('simulated data'), we ran several common clustering algorithms over a large range of k for each dataset, and evaluated each clustering run with several cluster optimality metrics. Clustering algorithms that repeatedly produced clusters with 3 or fewer members, and metrics that repeatedly produced monotonic results, were thrown out. We then checked whether or not the metrics tended to give the 'correct' cluster number in the simulated data, and threw out metrics with frequent incorrect optima. Finally, the remaining metrics underwent correlation analysis, and representative metrics from each correlated cluster were chosen as the 'final' metrics. The set of optimized clustering algorithms and metrics were then run over the input datasets with sequentially more variables added in.

Results

The mean normalized metric results were plotted over the range of included variables and the range of k to produce a 3D map of cluster optimality. This process was repeated for generated Gaussian data with known k, as well as all cancers in the TCGA with microarray

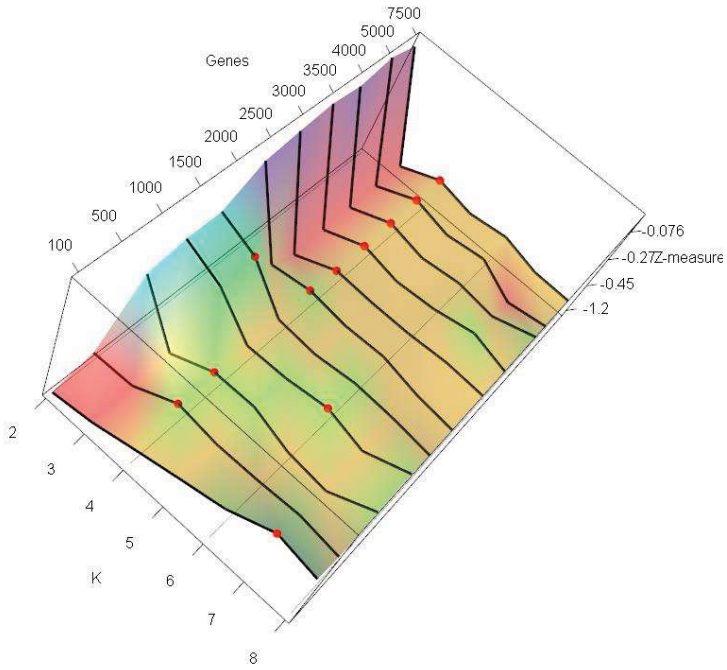


Figure 1. COMMUNAL clustering result on the TCGA rectal cancer mRNA dataset. X-axis: k; Y axis: number of genes included; Z-axis: normalized combined validity measures.

gene expression data. The simulated data were called correctly, even in the presence of 1:1 informative:noisy variables. The COMMUNAL optimal k matched TCGA optimal k for NMF consensus clustering for glioblastoma, rectal cancer, and uterine cancer, but revealed new insights about stability of cluster number assignment for all cancers. The output for the rectal cancer mRNA dataset is shown in Figure 1.

Conclusions

Different clustering algorithms and optimality metrics yield may different results for the same data. With COMMUNAL, we combine the signals from multiple different clustering algorithms and metrics to improve the signal:noise ratio and increase the robustness of a given chosen k. Furthermore, by mapping these calls over a gradually increasing number of variables for a given dataset, the stability of a clustering call to additional information can be assessed. COMMUNAL may be a useful tool for determining k in complex datasets, and is now available as an R package on CRAN.

Can automotive Big Data Analytics be applied for Intensive Care of Neonates? The issue of time.

Sidhartha Tan, MD and KP Unnikrishnan, PhD,

NorthShore University HealthSystem, Evanston, IL, USA

Summary: Big Data in the automobile industry has analogies to Big Data in a neonatal intensive care unit (NICU). Frequent examinations of data ensure the early detection and management of clinical problems. Big-Data is grossly underutilized in NICU. We present the challenges in transitioning to a Big-Data oriented NICU care.

Introduction: Modern-day automotive manufacturing and the smooth running of an automobile are sophisticated processes, managing many factors and variables. Big Data analytics solves many problems in different milieus of both areas. The analogy to medical care of a baby in a neonatal intensive care unit (NICU) is uncanny. A newborn baby is like a new car. Frequent examinations of data from sensors on a baby ensure the early detection of clinical problems and their management in the NICU. The driver and mechanic are analogous to the caretakers of sick babies in the NICU. There are expectations of a good outcome in both automotive industry and NICU. However, Big Data that is routinely utilized in the manufacturing and care of automobiles is grossly underutilized in NICU. There is a paucity of real-time data gathering, real-time data analysis, and on-the-fly adjustments based on Big Data. Our hypothesis is that NICU care can benefit from lessons learnt in the automotive industry. Implementation of analogous improvements will better clinical outcome of patients, improve clinical acumen of caretakers and provide new and exciting areas of research.

Methods: Previously, the vital sign data from the NICU was sent to a Philips server in proprietary format and then converted to HL7 format once a minute. This was then entered into to the electronic medical record when the nurses clicked on Epic forms once-an-hour or so. From the EMR, this irregularly sampled data was pulled out into an enterprise data warehouse (EDW) at the end of the day. The original HL7 data was kept initially only for 24 hours. We first created a real-time EDW engine to receive and store all data at one-minute intervals. We then had to deal with the issues of time in the NICU. In order to change the mindset from assessing the patient episodically to assessing in real time, observations had to be accurately recorded with respect to an actual clock and special Epic forms built for this purpose.

Results: In dealing with time, the typical nurse's role was re-assessed. We had to change our mindset from looking at a typical NICU patient at periodical intervals to assessing real time data. The culture of healthcare is episodic care based on periodic evaluations, much different from real time analytics, processing and control as in the automotive industry. The culture in the NICU needed to change from evaluating data at 4-hourly, hourly, and minute intervals to finer precision, i.e. at second and millisecond frequencies. The data collection stream had to be separated from the EMR data stream to store data at minute precision. We found that most observations in the EMR were not accurate to the minute. Finally, to convince nurses of the utility of taking extra care of time, we needed to provide a reason for taking the extra effort and engage the nurses in the cultural change.

Discussion: The challenges come not only from algorithms, software, and hardware, but also from the mind-set of immediate caretakers, supervisors, support staff, and administration. The biggest hurdle to overcome is the culture of overall medical care. The biggest challenge a more Big Data oriented care in NICUs is the issue of accurate time.

Improving Lupus Phenotyping Using Natural Language Processing and Machine Learning

Clayton Turner[1], Paul Anderson, PhD[1], James C. Oates, MD[2], Diane L. Kamen, MD[2], Jihad S. Obeid, MD[2]
[1]College of Charleston, Charleston, SC, [2]Medical University of South Carolina, Charleston, SC

Abstract: *Clinical notes from electronic health records are analyzed in order to determine patients' diagnosis of SLE. We used the cTAKES/*YTEX *natural language pipeline, and applied machine learning in an attempt to improve classification of patients with SLE. We plan to improve accuracy by using deep learning techniques.*

Introduction: A good portion of the information in the Electronic Health Record (EHR) is trapped in clinical notes that are not computable. Moreover, ICD-9 billing codes, although useful for phenotyping, are often misleading due to coding errors and/or billing criteria. Natural language processing (NLP) pipelines have proven successful in converting free-form text into grammar-less datasets, which conform to the needs of machine learning (ML). The goal of this research is to improve the accuracy of EHR phenotyping for Systemic Lupus Erythematosus (SLE), which is currently highly dependent on ICD-9 codes, by creating a classifier using an NLP/ML pipeline for clinical notes. Automated phenotyping is critical for identifying cohorts for clinical trials, genotype-phenotype studies and epidemiological disease burden studies.

Methods: With IRB approval we created a database that includes the clinical notes for patients (n=5047) seen in our rheumatology clinic at the Medical University of South Carolina (MUSC) over the period of one year (2010). The database included the SLE ICD-9 code status for all patients. Each patient typically had several clinical notes. Of these patients 395 had a gold-standard label of SLE status (1=yes, 0=no) determined by a team of rheumatologists. The Apache clinical Text Analysis and Knowledge Extraction System (cTAKES) pipeline, a natural language processing engine, takes unstructured clinical notes and outputs a descriptor of all of the UMLS concepts ids (CUIs) or bag of CUIs for each of the clinical notes (Savova 2010). For example, words like "history", referring to a patient history, and phrases like "malar rash" would be flagged as concepts. However, cTAKES does not produce output for machine learning algorithms. The Yale cTAKES Extensions (YTEX) adds that functionality as well as Negex for negation detection (Garla 2011). We applied the YTEX pipeline to our patient database to produce sparse Attribute-Relation File Format (ARFF) files of CUIs to be used as features for the ML algorithms. Our initial investigation was carried out using the Waikato Environment for Knowledge Analysis (WEKA) (Garner 1994, Hall 2009). Ten-fold cross validation classification was done in WEKA to classify SLE status using the following classifiers: ZeroR, Naive Bayes, Decision Tree, and Random Forests. Performance was compared to using ICD-9 codes for SLE (710.0) alone as a classifier.

Results: Accuracy and Area Under the Curve (AUC) for SLE on the 395 gold-standard samples were as follows: ZeroR 78.43% with AUC of 0.5, Naive Bayes 77.92% with AUC of 0.74, Decision Tree 93.66% with AUC of 0.93, and Random Forests 89.85% with AUC of 0.86. Using only the ICD-9 codes as a predictor in a decision tree for whether or not a patient has SLE yielded an accuracy of 97.45% with AUC of .925 when a patient was assigned the ICD-9 code of 710.0 at seven or more instances throughout their EHR timespan.

Discussion: The Naive Bayes algorithm assumes that the features used to classify the diagnosis are all independent, a requirement which is violated when diagnosing SLE, since diagnosis is based on the presence of 4 of 11 American College of Rheumatology (ACR) criteria. The decision tree and random forests algorithms both operate by selecting features which are most important to classifying a patient's diagnosis. For example, if the concept "Systemic Lupus Erythematosus" shows up in a patient's note multiple times, then that patient most likely will be flagged as having SLE since the concept of the disease name itself is indicative of a diagnosis. Additionally, SLE symptoms or ACR criteria, such as malar rash, tend to be important features for classification. Using ICD-9 alone for prediction appears to yield the highest accuracy, however it is not a statistical improvement over the decision tree built with NLP ($\alpha = 0.05$). The conclusions based on these results are limited by the small sample size (395 labeled samples, only 80 samples labeled as no SLE). We hypothesize that increasing the sample size to captures more borderline cases and more cases in general will increase the performance of NLP/ML prediction methods and not improve the ICD-9 based predictions. In addition to expanding the sample size, we will investigate semi-supervised methods of classification (e.g., deep learning) that leverage the available large unlabeled pool of samples. Finally, we will evaluate whether the results improve with feature selection and unsupervised learning algorithms, such as sparse coding and independent component analysis.

Conclusion: The pipeline which is being developed utilizes state-of-the-art NLP tools and ML techniques in order to predict whether or not a patient has SLE from free-form clinical notes. We plan to improve this pipeline by applying more cutting edge feature selection and deep learning algorithms.

Acknowledgements: This work is funded by the National Institutes of Health (NIH) Grant #s P60AR062755 and UL1TR000062, the Medical University of South Carolina, the College of Charleston and the SmartState Program in SC.

Challenges of cross-platform searches in time series microarray data

Guenter Tusch, PhD, Olvi Tole, MS, Mary Ellen Hoinski, MS
Medical and Bioinformatics Graduate Program,
Grand Valley State University, Grand Rapids, MI

Abstract

While microarray technology is still very important in molecular biology, there is also a large body of information available through a growing number of studies in public repositories like NCBI GEO and ArrayExpress. Software is now available to allow for cross-platform comparison. Temporal translational research is based on e.g. stimulus response studies, and includes searching for particular time pattern like peaks in a set of given genes across studies and platforms. Such a tool is not available. This study explores the feasibility based on our SPOT software.

Introduction and Background

Today microarray technology is still an important technology to assess gene expression in molecular biology. There is a wealth of information available, for instance, the US National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) has data of more than 1.2 Mio samples, ArrayExpress from the European Molecular Biology Laboratory has more than 1.5 Mio assays. A growing number of those are now studies including time series data based on stimulus response studies. Often peaks in gene profiles (identified, e.g., by a significant average change from one time point to the next) in temporal microarray studies represent a biological effect that is reversed after some time. In temporal translational research a researcher typically obtains a fold change profile and tries to retrieve similar profiles in a set of genes or gene products in microarray databases or clinical databases (that more frequently include microarray data, whole-genome sequencing, or other next-generation sequencing data).

Assume the researcher conducted a temporal study where she discovered peaks in a set of genes might be found in the same pathway. She now wants to see if finding the same effect in related studies can extend her hypothesis. Although different studies address similar questions a comparative search through the database is impeded by the use of heterogeneous microarray platforms and analysis methods. Researchers who perform high throughput gene expression assays often deposit their data in public databases, although the heterogeneity of platforms leads to challenges for the combination and search of these data sets. Earlier publications have suggested that some of the variability in cross-platform studies was due to annotation problems that made it difficult to reconcile which genes were measured by specific probes. These issues have been resolved recently (e.g., [1]). The problems can be overcome by carefully selecting high quality datasets, as has been done for certain problems on the INMEX website (http://www.inmex.ca). A challenge is cross platform normalization where nine different methods are currently available, and no rigorous comparison exists. Furthermore, software for the selected method must be obtained and incorporated into a data analysis workflow.

Methodology

We explore how knowledge-based temporal abstraction[2], where time-stamped data points are transformed into an interval-based representation. Each interval represents a specific trend, e.g., “increasing”, “decreasing”, or “constant”, defined by statistical significance. A “peak” can be defined as an increasing interval immediately followed by a decreasing one. Thus peaks can be found even if not all experiments use the same time points. If as in our example a researcher tries to extend the finding in a local dataset by searching for similar public datasets "similarity" would mean that a peak could be found in a specified interval. Because it is based on significance, the sample sizes of the different studies factor in as well and a non-match could simply occur due to a small sample size.

System Implementation and Evaluation

For implementation we utilized a software platform SPOT[1] based on open-source software, R and Bioconductor. We evaluated our approach a wide array of temporal studies from NCBI GEO.

References

1. MAQC Consortium, Nat Biotechnol. 2006 September ; 24(9): 1151–1161.
2. Tusch G, Bretl C, O'Connor M, Das A, SPOT--towards temporal data mining in medicine and bioinformatics, AMIA Annu Symp Proc. 2008: 1157.

DRUGSDB: Annotating Medicines, Diseases and Targets

Oleg Ursu and Tudor I. Oprea
Translational Informatics Division, Department of Internal Medicine
UNM School of Medicine, Albuquerque NM 87131, USA

Introduction: Open-access drug information sites like Dailymed index approved drug labels (ADLs), whereas chemically-cognizant platforms like Drugbank focus on active pharmaceutical ingredients (APIs). While each of these resources annotates medicines, there is no direct link between ADLs to APIs. Furthermore, few of these resources follow vocabularies used by the FDA or by the Observational Health Data Sciences and Informatics (OHDSI) program. For now, text- or chemical structure based- queries are not possible on the DailyMed platform; and OHDSI-compliant vocabularies for indications and contra-indications cannot be queried in any of the above resources. At University of New Mexico, we are developing DRUGSDB, a repository for therapeutic agents that aggregates information on approved and discontinued drugs worldwide, with focus on small molecules, which maps APIs to ADLs and vice-versa.

Methods: Chemicals: chemical structures, FDA chemical nature (e.g., organic, inorganic), World Health Organization (WHO) INN (International Nonproprietary Names) stem and ATC (Anatomic Therapeutic Chemical) codes were curated for each API using controlled vocabularies and precise chemical structure representation rules. Specific rules were observed regarding targets (e.g., UniProt recommended names and identifiers), bioactivities (all converted to the negative log10 of the molar concentration) and bioactivity types. Drugs: ADLs extracted from DailyMed, in full compliance with current FDA annotations and definitions, were cross-matched against the Orange Book, via LOINC (Logical Observation Identifiers Names and Codes) sections headings. Indications, contra-indications and off-label indications were indexed using the OHDSI, UMLS compliant vocabulary. We mapped ADLs and APIs to National Drug Codes (NDC), compatible with RxNorm.

Results: As of September 23, 2014, DRUGSDB maps 50,311 ADLs onto 1608 APIs (of 4290 total). 2747 APIs are annotated with ATC codes (4613 codes, total), and 2354 are annotated onto 521 unique INN stems. 1944 APIs are mapped onto 2760 unique protein targets (1906 human), with 13,803 human & 2713 non-human numeric bioactivity values (over 27,000 annotations). DRUGSDB has 3,372 disease concepts (2,412 as indications, with another 331 off-label indications) associated with 1,858 APIs, as mined from ADL data.

Discussion: When comparing human prescription (1343 APIs) vs. over-the-counter (OTC) medicines (265 APIs), we found more OTC drugs (27,109 ADLs) compared to prescription (23,202 ADLs), respectively. Many ADLs contain the same APIs: For example, 3419 ADLs contain acetaminophen, though only 84 fixed dose combinations are approved for this API. DRUGSDB maps 2520 mechanisms of action (MoAs) onto 1241 APIs; of these, 1448 are linked to numeric values. As bioactivities range from micromolar to (sub) nanomolar, it becomes evident that affinity is not the only factor driving efficacy. Target co-localization, tissue bio-concentration and efflux pumps are likely to play an equally important role, not only for efficacy but also for safety. DRUGSDB supports computational workflows for drug repurposing – for example by linking off-label indications to targets for nearly 700 APIs, as well as “on”-label and contraindications. As we link diseases to targets, we seek clinical relevance for drug-influenced biochemical and pharmacological events, in order to streamline the process of computer-aided drug repurposing.

Development of a Power Analysis Tool for Biomarker Discovery Studies

Garrick Wallstrom, PhD, Michelle Winerip, BA, Michael Fiacco, BS, Joshua LaBaer, MD, PhD
Center for Personalized Diagnostics, The Biodesign Institute
Arizona State University, Tempe, AZ

Summary
We are developing a web-based power analysis tool to assist investigators that are designing biomarker discovery studies. The tool will accommodate multiple experimental designs, commonly used analytic strategies and will account for disease heterogeneity.

Introduction and Background
Despite the large number of biomarker discovery studies there has been scant attention paid to sample size requirements for such studies. The very few publicly available tools that exist today that allow researchers to conduct power analyses for biomarker discovery studies are severely limited by their scope and accessibility to researchers. Further, these analyses are complex and typically cannot be performed using textbook formulae. Instead, they require specialized expertise in the statistical methods of biomarker research. The lack of power analysis tools for biomarker studies and the complexity of such power analyses suggest that most biomarker studies are undertaken without proper examination of statistical power. An underpowered study may fail to find biomarkers due to inadequate sample sizes while an overpowered study wastes precious resources.

Most studies that have examined sample size requirements for biomarker studies or the relative performance of statistical biomarker selection methods have not considered these questions in the context of heterogeneous diseases. However we have recently shown that the required sample sizes for a heterogeneous disease are more than twice than for a homogeneous disease (1). Further, the predominant methods used ubiquitously for biomarker studies, including the full area under the curve (AUC) evaluation of a receiver operator characteristic (ROC) curve and t-testing, performed extremely poorly for heterogeneous diseases.

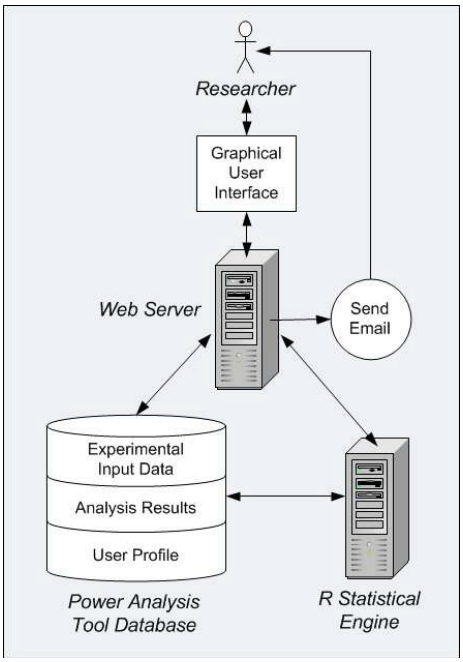


Figure 1: Architecture of the web-based power analysis tool for biomarker studies.

Methods
The web-based biomarker discovery power analysis tool will be designed and developed to help users determine an effective experimental design and sample size requirements for their biomarker study, and guide the analytic strategy for their data. The architecture of the tool is shown in Figure 1. The primary target users are statisticians involved in the planning or analysis of biomarker studies. However the tool will also be accessible to statistically knowledgeable biomarker researchers. Although the computational engine behind the analyses is R, users will not need to know R programming to use the web-based tool.

Results
We are currently developing the web-based power analysis tool. The anticipated launch date is March 1, 2015.

Conclusions
This tool will enable researchers to determine sample size requirements and examine multiple analytic methods for discovery research while properly accounting for disease heterogeneity.

References
1. Wallstrom G, Anderson KS, LaBaer J. Biomarker discovery for heterogeneous diseases, Cancer Epidemiology Biomarkers and Prevention. 2013 May. 22(5): 747-55.

New Evaluation Metric for Protein-protein Interaction Prediction

Haohan Wang[1], Madhavi K. Ganapathiraju, PhD[2]
[1]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA;

[2]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

Abstract

We propose an evaluation metric focusing on the eligibility of a machine learning model to be adopted by biologists. This metric evaluates whether a predictor can predict only new interactomes with high precision, meanwhile, it eliminates the influence of test dataset, allowing comparisons across data sets.

Introduction

Machine learning based identifying protein-protein interaction pairs is a fundamental step for investigating protein functions. Most of these machine learning algorithms are proved to be promising with one or several traditional evaluation metrics. However, in order for the predicted interactions to be directly adopted by biologists, who perform expensive experiments, the predictions have to be of high precision, even if the recall is low. With only traditional metrics like accuracy, ROC, or precision-recall curve etc. this aspect cannot be evaluated or numerically represented fully, without impact of the distribution of testing data set. Therefore, we propose this metric.

Methodology

With a weak, yet realistic, assumption that machine learning algorithm must predict at least one positive interactome, we propose an evaluation metric lives in a 3D dimension space from the two most famous criteria, ROC and precision recall curve. Instead of working on traditional precision, which is defined as $tp/tp+fp$, we only focus on the purity of predicted interactomes, i.e. ratio of true interactomes predicted and non-interactomes predicted, (tp/fp). The essential difference between these two is that precision is lower bounded by the portion of positive data in test data set (bound is reached when we set a threshold to predict everything to be positive), while our ratio falls into the closed interval between 0 and infinity. In addition to the first metric where we exaggerate on the purity of predicted positive interactomes, we also focus on describing the numbers of false positive interactomes predicted. The second dimension of our metric is described as $1\text{-}fp/tn$. We add a minus sign to simplify the explanation of numeric values (larger values will mean better performance). The third dimension serve as the connection between these two dimensions, we choose recall here. Traditional recall can connect these two dimensions and it can also capture the fact that for the classifiers that are evaluated equivalent in the ability of predicting new interactomes, the one that is able predict more interactomes should be evaluated better. One drawback our this metric is the lack of upper bound, thus the result cannot be visualized.

Experimental Result

To validate our proposed method, we evaluate the performance of our evaluation metric in comparison of the others. Limited by length of this abstract, we only show the results compared with precision recall curve and ROC. The experiment is performed with a series of simple predictors, rather than real world models because we are able to control the predictors this way. We first evaluate the same classifier with data sets of different distribution over positive and negative instances. ROC curve and our metric produces the same result, PR curve evaluates results that are correlated with the positive portion of data. Then three experiments are carried out with: 1) classifiers with increasing ability of predicting interactomes; 2) classifiers with decreasing ability of excluding non-intreactomes 3) classifiers combined of 1) and 2). Results show that ROC curve only focuses on the general performance of a classifier, regardless of the differences of positive and negative data. Precision recall curve and our metric favors the classifier that can predict positive interactomes.

Discussion

We solved the problem that current evaluation metrics cannot evaluate the performance of a model when adopted by biologists, where high precision is required, with a universal result that is irrelevant with test data sets.

A Novel Algorithm for Automated Data Extraction from Free Text Pathology Reports in Patients Undergoing Adrenal Surgery

Qun Xiang, MD, Jennifer Rabaglia, MD

From the Division of Gastrointestinal and Endocrine Surgery

Department of Surgery, the University of Texas Southwestern Medical Center

Dallas, TX

Background: Pathology reports housed in electronic medical records store critical information regarding patients and their various tissue diagnoses. They are a key component in the evaluation a vast majority of patients with adrenal tumors. Currently, pathology is reported using free text, making the data extraction process labor and time intensive for providers and researchers who wish to review large numbers of records. Automation of this process could save significant time and cost when dealing with research and patient registries. Some in informatics have concentrated on classifying text to determine whether tumors are called benign or malignant within path reports. Others have focused on deeper parsing of entity relation. Our aim is to develop a natural language processing (NLP) algorithm to automate relevant data extraction from pathology reports, in order to provide fast, complete and accurate pathology data extraction for clinicians and researchers.

Material and Methods: All patients who underwent adrenal surgery for tumors at a single center between 2006 and the present were included in this study. Of these, 203 cases with associated pathology reports were extracted from the EPIC Clarity server. We planned to automatically extract key information from reports, including side of adrenal tumor, tumor size, pathology diagnosis, lympho-vascular invasion, and metastasis to lymph nodes or other sites. A regular expression based algorithm was developed on a training set of 148 reports. Each report was divided into sentences. Part-of-speech tagging is used to tag each word of one sentence. Then words were chopped into different phrases,

noun phrase (NP), verb phrase (VP) and prepositional phrase (PP). Lexicons are defined for diagnosis, tumor, lesion, and cyst. Rules were written to find relationships for different fields. For example, NoduleCD, a rule use to catch up the size of a nodule, is written like this: '<lesion|nodule><VP><PP>?<MCD><TO>?<MCD>?'. When correct relationship was found, corresponding value was extracted.

Python 2.6 NLTK package was used to build the application for extracting information from pathology reports. An additional 55 reports were used as the testing data set. One trained data abstractor (DX) manually confirmed the correct value for each data field, and this was compared against the automated testing data set results. Precision, Recall, F1 score and Accuracy were calculated by using MS Excel 2010.

Results: Among the 60 reports in the testing data set, there were a total of 360 individual data points (results). Of these, the NLP algorithm made a total of 17 mistakes in 15 subjects. After running statistical analysis on the result (Table 1), we reached Precision (80% - 100%), Recall (87.5% - 100%), F1 Score (83.6% - 100%), Accuracy (78% - 100%).

Conclusion: This is the first known NLP algorithm designed for automated relationship extraction from adrenal surgical pathology reports. The algorithm has high Precision, Recall, F1 Score and Accuracy. Most of the mistakes were related to the size of tumor, likely due to the complicated free text grammar structure utilized when referring to size in 3 dimensions. We hope to refine this machine learning algorithm in order to improve accuracy of the data extraction in the future. However, it is clear that NLP algorithms such as this which facilitate rapid, accurate, automated data extraction from free text reports may serve as a revolutionary tool for clinicians and researchers aiming to evaluate and learn from large patient cohorts. This study provides a solid foundation for the development of an automated system for ascertainment of outcomes and the potential for valuable real time feedback to clinicians and researchers.

Table 1: Precision, Recall, F1 Score and Accuracy regarding different prospects of adrenal tumors

	Precision	Recall	F1 Score	Accuracy
Size	0.825	0.892	0.857	0.78
Side	1	1	1	1
Diagnosis	0.981	0.963	0.972	0.946
Lymphvas - Invasion	1	0.929	0.963	0.929
Lymph Node	1	1	1	1
Metastasis	1	1	1	1

Are All Vaccines Created Equal? Using Electronic Health Records to Discover Vaccines Associated With Clinician-Coded Adverse Events

Mary Regina Boland, MA[1,4] and Nicholas P Tatonetti, PhD[1-4]

[1]Department of Biomedical Informatics, [2]Department of Medicine, [3]Department of Systems Biology, [4]Observational Health Data Sciences and Informatics, Columbia University

Abstract

Adverse drug events (ADEs) are responsible for unnecessary patient deaths making them a major public health issue. Literature estimates 1% of ADEs recorded in Electronic Health Records (EHRs) are reported to federal databases making EHRs a vital source of ADE-related information. Using Columbia University Medical Center (CUMC)'s EHRs, we developed an algorithm to mine for vaccine-related ADEs occurring within 3 months of vaccination. In phase one, we measured the association between vaccinated patients with an ADE (cases) against those vaccinated without an ADE. To adjust for healthcare-process effects, phase two compared cases against those who returned to CUMC within 3 months without an ADE. We report 7 results passing multiplicity correction after demographic confounder adjustment. We observed an association, having some literature support, between swine flu vaccination and ADEs (H1N1v-like, OR=9.469, p<0.001; H1N1/H3N2, OR=3.207, p<0.001). Our algorithm could inform clinicians of the risks/benefits of vaccinations towards improving clinical care.

1. Introduction

1.1 Adverse Drug Events Are Important for Public Health

Adverse drug events (ADEs) are a major cause of death in the United States of America [1]. To address this serious public health issue, the Federal Drug Administration (FDA) developed an adverse event reporting system. Since this reporting system began, more than 75 drugs or drug products have been removed from public use [2]. The number of ADEs occurring between 1998 and 2005 increased 2.6 fold illustrating the increasing importance of ADE prevention in clinical care [3]. Many ADE detection methods rely on adequate physician, pharmacist, or nurse reporting of the ADE to federal reporting systems. Realizing that 1% of ADEs recorded in Electronic Health Records (EHRs) are reported on the federal level [4], we chose to harness the large set of clinician-reported ADEs available in EHRs to find novel vaccine-ADE associations.

1.2 Informatics Methods Enable Harnessing of Data Within EHRs

The widespread adoption of EHRs enables meaningful use [5] of data recorded during the clinical encounter. Appropriate use of EHR data requires overcoming definition discrepancies [6], data sparseness and quality [7], bias [8], and healthcare process effects [9]. Informatics methods overcome these challenges by employing standardized ontologies to minimize definition discrepancies [10-12], measuring concordance across integrated datasets for data sparseness and quality assessment [7], and minimizing bias and healthcare process effects using statistical methods [13]. Informatics methods applied to EHRs [14] have been successful in diverse areas [15-17] including pharmacovigilance [18, 19]. They are also useful in predicting ADEs using chemical and molecular structures of compounds [20]. Approaching the problem from a different angle, our method investigates ADEs occurring and recorded during routine clinical care.

1.3 ADE Detection and Prevention Feasible Using EHRs

Multiple algorithms have shown the usefulness of EHRs for ADE detection. Haerian et al. developed a method for identifying drugs associated with two serious ADEs, rhabdomyolysis and agranulocytosis, after adjusting for patient comorbidities [19]. Luo et al. developed a pattern mining method for detecting ADEs from clinical trials data [21]. Linder et al. found that only 1% of EHR recorded ADEs are reported to the federal government, demonstrating that EHRs are a rich data source for ADE detection [4].

2. Materials and Methods

2.1 Columbia University Medical Center Dataset

We used EHR data from Columbia University Medical Center (CUMC), previously converted to the Common Data Model (CDM) [22] developed by the Observational Medical Outcomes Partnership (OMOP). This dataset contains patients' drug-related and diagnosis information. The CUMC Institutional Review Board approved this study.

2.2 An Algorithm to Mine for Vaccines Associated with Adverse Events

We mapped all *International Classification of Diseases, version 9* (ICD-9) codes to the *Systemized Nomenclature of Medicine – Clinical Terms* (SNOMED-CT) using the OMOP CDM v.4 [22], which was proven useful by a number of prior research studies [23, 24]. By taking advantage of the medication-terminology mapping in the CDM (which

includes both RxNORM and NDF-RT)[22] we are able to map many different vaccines from different manufacturers to the same core ingredient set. Others obtained high quality results when using this same CDM mapping for medications [24]. Using the CDM also helps minimize terminology mapping issues common when using EHRs for medication information [25].

In the OMOP CDM [22], one code for "adverse effect due to correct medicinal substance properly administered" maps to 75 ICD-9 codes (each with relatively low prevalence). We used this mapping and extracted a population of 16,296 patients with a coded ADE from an appropriately prescribed and administered drug. Because we were interested in vaccines, we extracted all patients who were vaccinated in our medical system (N=70,050). Subsequently, we recorded patients as having a vaccine-related ADE if the ADE occurred within a 3-month window (i.e., 90 days) after the vaccination date. We selected a 3-month window because there is literature suggesting that over 8 weeks time may be necessary to appropriately capture a vaccine-related ADE [26]. If several ADEs occurred within the 3-month time frame (e.g., one 2 days, and another 7 days after vaccination) then both were included in the analysis. This was done because both clinician-coded ADEs could be the result of the vaccination.

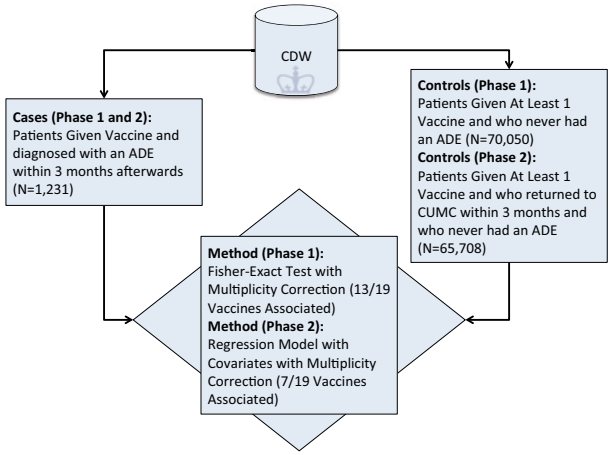


Figure 1. Algorithm Schema to Detect Vaccines Associated with Clinician-Coded Adverse Events

2.2.1 Phase One: Mining Vaccine-ADE Associations Across All Vaccinated Patients

The first part of our algorithm (**Figure 1)** calculates the association between each vaccine and an ADE within 3 months by comparing each individual vaccine (case) to all other vaccines in our dataset (as controls). Controls include all patients who were vaccinated regardless of whether they returned to the hospital for a follow-up visit. Associations are measured using the fisher-exact test with multiplicity correction using Bonferroni's method (R v.3.1.0).

2.2.2 Phase Two: Mining Vaccine-ADE Associations Adjusting for Health-Care Process and Demographic Effects

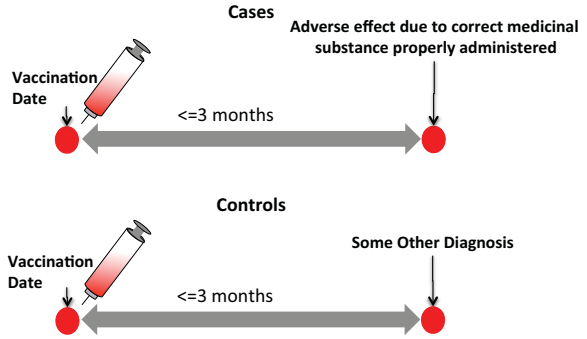


Figure 2. In Phase Two, Controls were Selected that Returned to CUMC within 3 Months Minimizing Healthcare Process Biases that Affect Patients' Ability to Return for Treatment.

To adjust for various health-care process effects [8, 9] that may affect whether or not a patient returns to CUMC within 3 months, we decided to use as controls all patients who were vaccinated and were subsequently diagnosed with some other medical condition (not an ADE) within 3 months of the vaccination date. Our cases remained unchanged and consisted of all vaccinated patients with an ADE diagnosis within 3 months. Therefore, in this second phase of the algorithm both cases and controls returned to CUMC within 3 months. For this analysis, we had 65,708 controls and the same 1,231 cases (**Figure 1)**. We measured the association between each vaccination and an ADE diagnosis using logistic regression. Specifically, each potential confounder (i.e., ethnicity, race, sex, age (at time of vaccination)) was modeled as a covariate in the logistic regression equation with the binary response (outcome) variable indicating the presence or absence of an ADE within 3 months of vaccination and the predictor variable denoting presence or absence of the vaccine of interest (R v.3.1.0). An association is reported as significant if the Bonferroni adjusted p-value is <=0.05. We further illustrate phase two's control selection method in **Figure 2.**

3. Results

3.1 Overview of CUMC Dataset

Our dataset contained 472,451 patients with both medication and diagnosis-related information. We found 19 vaccines prescribed at CUMC with at least one patient with a recorded ADE within 3 months after vaccination. In total, 1,231 vaccinated patients were diagnosed with an ADE within 3 months, and **Figure 3** depicts their characteristics.

3.2 Vaccine-ADE Algorithm

3.2.1 Phase One: Mining for Associations Across All Vaccinated Patients

We applied our algorithm to all 19 vaccines and measured the association between the vaccine's administration and an ADE within 3 months afterwards. After Bonferroni adjustment, we found 13 vaccines were associated with an ADE at this step. Characteristics of our case patients are shown in **Figure 3** (note that the case population did not change between methods). All results from both phases of the algorithm are provided in **Table 1** (following page).

3.2.2 Phase Two: Mining for Associations After Adjusting for Healthcare Process, and Demographics

During phase two, we constructed a logistic regression model with covariates for age, ethnicity, race and sex. We report results passing multiplicity correction (7 of 19 vaccines). Four vaccines were significantly associated with more ADEs and three vaccines were significantly associated with fewer ADEs compared with other vaccines. Two of the four vaccines associated with more ADEs were vaccines against flus originating in swine including: H1N1/H3N2/inactivated B-Brisbane-60-2008 strain and H1N1v-like virus vaccine. For the H1N1/H3N2 combo vaccine, 503 of 6,904 patients returning within 3 months experienced an ADE. Further, for the H1N1v-like vaccine, 103 of 423 patients returning within 3 months experienced an ADE. Both were significant after adjusting for demographic confounders (adjusted p<0.001 for both, **Table 1).** In several instances we found that all patients who returned to CUMC within 3 months had an ADE diagnosis. This includes five vaccines typically given to infants: pertussis/diphtheria/tetanus; hepatitis B surface antigen; tetanus; diphtheria/haemophilus B; Polio 3 types. While interesting, none of these vaccines were significant after adjusting for demographic confounders and multiplicity.

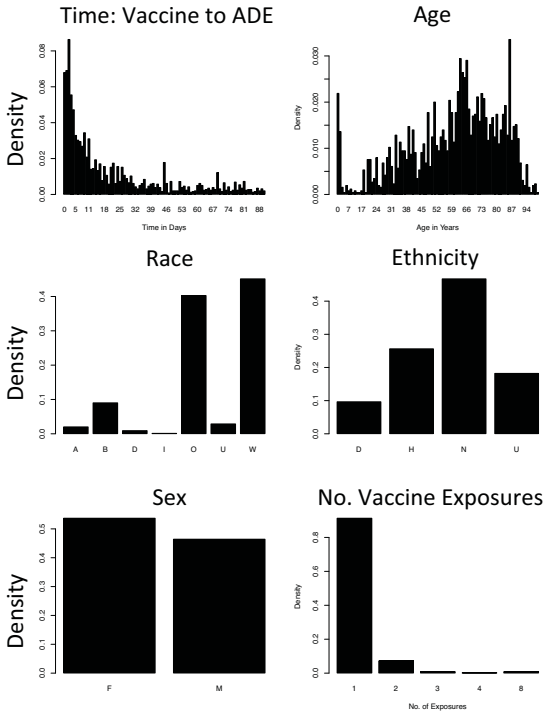


Figure 3. Characteristics of Patients That Developed an ADE within 3 months after Vaccination

4. Discussion

4.1 Important Vaccine-ADE Associations

Vaccine-related ADEs can result from a number of different mechanisms important for achieving precision medicine [27]. Our two-phase algorithm was developed specifically for finding vaccines associated with clinician-coded ADEs in EHRs and was agnostic to the mechanism underlying the vaccine-ADE relationship. Interestingly, two types of swine flu vaccines were positively associated with an increased risk of an ADE within 3 months of vaccination after adjustment for confounders (**Table 1),** namely the combo H1N1 / H3N2 / B-brisbane influenza (OR=3.207, p<0.001) vaccine and the influenza A-California-7-2009-(H1N1)v-like virus (OR=9.469, p<0.001) vaccine. Importantly, H1N1 originates in swine [28] and all swine flu vaccines in our study were associated with increased risk of ADEs. This fits well with prior literature supporting vaccine-related ADEs resulting from a different swine flu vaccine in the 1970s [26], which resulted in very serious ADEs including paralysis. Another study, found a similar result for H1N1 vaccination when compared to general influenza vaccination [29].

4.2 Value of Clinician-Coded ADE Associations

Early detection of ADEs is crucial for patient safety. Using our algorithm, we uncovered several vaccines that resulted in ADEs within 3 months for all patients who returned to CUMC within 3 months (**Table 1).** This was true for several vaccines given to infants. Although the results are not significant after covariate modeling (age is one confounder) it is suggestive of a relationship that may warrant further exploration. There are two main types of Hepatitis B vaccinations at CUMC: Hepatitis B (surface antigen) at a concentration of 0.04 mg/ml and Hepatitis B recombinant at a concentration of 0.01-0.02 mg/ml. Vaccination by the higher dose Hepatitis B vaccine resulted in 9 patients with ADEs out of 9 patients with the vaccine who returned to CUMC within 3 months (100% developed an ADE). Contrastingly, vaccination by the lower dose Hepatitis B recombinant vaccine (half to one-quarter the potency) resulted in ADEs among 1 of 1,617 patients seen at CUMC within 3 months after vaccination. Neither hepatitis vaccine was significantly associated with ADEs after adjusting for age, sex, ethnicity, and race. Patients receiving the higher dose Hepatitis B vaccine were less likely to return within 3 months (9/1474, **Table 1)** then those receiving the lower dose (1617/1659, **Table 1).** A likely explanation is that a higher proportion of infants received the lower dose (0.1-0.2 mg/ml) vaccine (94.45% of those vaccinated were <=0 years); whereas, both infants and toddlers received the higher dose (0.4 mg/ml) vaccination (65.94% of those vaccinated were <=0 years; 27.34% were one year olds).

Table 1. Vaccine-ADE Results for Phase 1 and 2 Trials.

			Phase 1: Association Between Vaccine Administration and Adverse Effect				Phase 2: Adjusting for Demographics and Only Including Patients Returning Within 3 Months			
Shortened Vaccine Name	**Origin Organism**	**No. Cases[1]**	**No. Vaccinated**	**Prop.[2]**	**Odds Ratio (OR)**	**Adj. P[4]**	**No. Vaccinated who Returned within 3 Months**	**Prop.[3]**	**OR**	**Adj. P[4]**
Associated with ADEs After Confounder Adjustment:										
Mumps	Human	5	785	0.006	0.451	1	8	0.625	106.793	$9.54X10^{-9}$
H1N1/ H3N2 / Inactivated B-Brisbane-60-2008 strain	Swine Virus (first 2), Human Virus	503	19517	0.026	2.119	$3.10X10^{-41}$	6904	0.073	3.207	$5.86X10^{-103}$
Pertussis / Diphtheria/ Haemophilus b / Polio / Tetanus	Bacteria (first 3), virus, bacteria	68	11575	0.006	0.399	$1.83X10^{-16}$	163	0.417	31.216	$9.42X10^{-90}$
H1N1v-like virus vaccine (0.25-0.5 mg/ml)	Swine Virus	103	3701	0.028	2.070	$1.90X10^{-9}$	423	0.243	9.469	$7.70X10^{-79}$
Associated with Fewer ADEs After Confounder Adjustment:										
Pneumococcal Type 1, 10A, 11A, 12F	Bacteria	1186	34983	0.034	4.149	$5.68X10^{-230}$	33976	0.035	0.406	$7.20X10^{-66}$
Rubella	Human	21	13568	0.002	0.101	$1.21X10^{-56}$	13565	0.002	0.055	$2.01X10^{-37}$
Pertussis / Diphtheria/ Hepatitis B Surface Antigen (0.02 mg/ml)	Bacteria (first 2), Primate virus	64	11406	0.006	0.381	$2.30X10^{-17}$	10660	0.006	0.208	$8.65X10^{-33}$
Insignificant After Confounder Adjustment:										
Hepatitis B (0.01 or 0.02 mg/ml)	Primate virus	1	1659	0.001	0.042	$5.37X10^{-8}$	1617	0.001	0.056	0.079
Varicella-Zoster Live (Oka-Merck)[Varivax]	Vertebrate Virus	3	99	0.030	2.208	1	30	0.1	5.618	0.103
Pertussis / Diphtheria / Tetanus [Infanrix]	Bacteria	59	11301	0.005	0.353	$3.35X10^{-19}$	59	1	8982647	1
Hepatitis B (0.04 mg/ml)	Primate virus	9	1474	0.006	0.431	0.133	9	1	135359422	1
Tetanus	Bacteria	57	11301	0.005	0.341	$4.28X10^{-20}$	57	1	89762087	1
Diphtheria / Haemophilus B	Bacteria	49	10242	0.005	0.325	$2.11X10^{-19}$	49	1	90639367	1
Polio Types 1-3	Virus	5	1453	0.003	0.242	0.002	5	1	82665283	1
Meningococcal Group A/C/W/Y	Bacteria	10	291	0.034	2.519	0.161	273	0.037	1.154	1
Measles / Mumps / Rubella	Human	21	13571	0.002	0.101	$1.23X10^{-56}$	21	1	129593962	1
Streptococcus Pneumonia	Bacteria	3	545	0.006	0.390	1	56	0.054	2.665	1
Haemophilus B	Bacteria	13	13	0.008	0.600	1	1545	1	104864930	1
Diphtheria / Tetanus	Bacteria	58	11301	0.005	0.347	$1.42X10^{-19}$	58	1	89078106	1

[1]Vaccinated and ADE within 3 months
[2]Cases / No. Vaccinated
[3]Cases / No. Vaccinated and Returned to CUMC Within 3 months
[4]Adjustment made using Bonferroni. Only Bonferroni-adjusted p-values <=0.05 were considered significant.

Infants have more wellness visits per year; therefore, vaccines given to a higher proportion of infants would be expected to have a higher return rate within 3 months (which we observed). This also demonstrates how the healthcare process can affect results of retrospective analyses using EHRs. Importantly, we adjusted for these types of biases in our algorithm by comparing patients receiving an ADE within 3 months to those who have returned to the hospital ADE-free within 3 months to help adjust for these biases. We also included age as a covariate in our regression model to adjust for age as well.

4.3 Limitations and Future Work

A limitation of our work includes our exclusive use of clinician recorded ADEs from EHRs. Some estimates suggest that only one-tenth of ADEs are clinician reported [30]. Therefore, we may be under-estimating the number of ADEs. We used only clinician-reported ADEs because we wanted to ensure that a clinician had validated the ADE as having occurred (i.e., a "true" ADE). Future work includes further exploration of dose-dependency effects for vaccine-related ADEs. Dosage data was only available for some vaccines at this stage. However, we hope to include clinical text and other data types in future to further tease out dosage effects and their relation to ADE risk.

5. Conclusion

We present an algorithm for discovering vaccines more likely to result in clinician-reported ADEs within 3 months of vaccination when compared to other vaccines. Our method found several interesting associations including two swine flu vaccinations that are positively associated with ADEs within 3 months of vaccination after confounder adjustment.

Acknowledgments: We thank George Hripcsak, MD for useful discussions on ADE coding in EHRs. Support provided by **T15 LM00707** and **R01 GM107145**. Authors report no conflicts of interest.

References

1. Lazarou J, Pomeranz B, Corey P. Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. JAMA. 1998;279(15):1200-5.
2. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the united states, 1969-2002: The importance of reporting suspected reactions. Archives of Internal Medicine. 2005;165(12):1363-9.
3. Moore TJ, Cohen MR, Furberg CD. SErious adverse drug events reported to the food and drug administration, 1998-2005. Archives of Internal Medicine. 2007;167(16):1752-9.
4. Linder JA, Haas JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, et al. Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. Pharmacoepidemiology and Drug Safety. 2010;19(12):1211-5.
5. Jha AK. Meaningful use of electronic health records: the road ahead. JAMA. 2010;304(15):1709-10.
6. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. J Am Med Inform Assoc. 2013;20(e2):e232-8.
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013;20(1):144-51.
8. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. J Biomed Discov Collab. 2011;6:48-52.
9. Hripcsak G, Albers DJ. Correlating electronic health record concepts with healthcare process events. J Am Med Inform Assoc. 2013;20(e2):e311-8.
10. Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the Content Coverage of SNOMED CT: Ability of SNOMED Clinical Terms to Represent Clinical Problem Lists. Mayo Clinic Proceedings. 2006;81(6):741-8.
11. Lin K, Hsieh A, Farzaneh S, Doan S, Kim H. Standardizing Phenotype Variables in the Database of Genotypes and phenotypes (dbGaP) based on Information Models. AMIA Jt Summits Transl Sci Proc 2013 2013;Mar 18:110.
12. Eilbeck K, Jacobs J, McGarvey S, Vinion C, Staes CJ, editors. Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC. ICBO; 2013.
13. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Using discordance to improve classification in narrative clinical databases: An application to community-acquired pneumonia. Computers in Biology and Medicine. 2007;37(3):296-304.
14. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012;13(6):395-405.
15. Boland MR, Hripcsak G, Albers DJ, Wei Y, Wilcox AB, Wei J, et al. Discovering medical conditions associated with periodontitis using linked electronic health records. J Clin Periodontol. 2013;40(5):474-82.
16. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics. 2010;26(9):1205-10.
17. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417-28.
18. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009;16(3):328-37.
19. Haerian K, Varn D, Vaidya S, Ena L, Chase H, Friedman C. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. Clinical Pharmacology & Therapeutics. 2012;92(2):228-34.
20. Liu M, Wu Y, Chen Y, Sun J, Zhao Z, Chen X-w, et al. Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. J Am Med Inform Assoc. 2012;19(e1):e28-e35.
21. Luo Z, Zhang G, Xu R. Mining patterns of adverse events using aggregated clinical trial results. AMIA Jt Summits Transl Sci Proc 2013. 2013;Mar 18:112-6.
22. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.
23. Boland MR, Tatonetti NP, Hripcsak G. CAESAR: a Classification Approach for Extracting Severity Automatically from Electronic Health Records. Intelligent Systems for Molecular Biology Phenotype Day. 2014;Boston, MA(In Press):1-8.
24. Ryan PB, Madigan D, Stang PE, Schuemie MJ, Hripcsak G. Medication-wide association studies. CPT: pharmacometrics & systems pharmacology. 2013;2:e76.
25. Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. J Am Med Inform Assoc. 2010;17(4):432-9.
26. Langmuir AD, Bregman DJ, Kurland LT, Nathanson N, Victor M. An Epidemiologic and Clinical Evaluation of Guillain-Barre Syndrome Reported in Association with the Administration of Swine Influenza Vaccines. American Journal of Epidemiology. 1984;119(6):841-79.
27. Peterson TA, Doughty E, Kann MG. Towards Precision Medicine: Advances in Computational Approaches for the Analysis of Human Variants. Journal of Molecular Biology. 2013;425(21):4047-63.
28. Trifonov V, Khiabanian H, Rabadan R. Geographic dependence, surveillance, and origins of the 2009 influenza A (H1N1) virus. New England Journal of Medicine. 2009;361(2):115-9.
29. Vellozzi C, Broder KR, Haber P, Guh A, Nguyen M, Cano M, et al. Adverse events following influenza A (H1N1) 2009 monovalent vaccines reported to the Vaccine Adverse Event Reporting System, United States, October 1, 2009–January 31, 2010. Vaccine. 2010;28(45):7248-55.
30. Classen DC, Resar R, Griffin F, Federico F, Frankel T, Kimmel N, et al. 'Global Trigger Tool' Shows That Adverse Events In Hospitals May Be Ten Times Greater Than Previously Measured. Health Affairs. 2011;30(4):581-9.

Disease Comorbidity Network Guides the Detection of Molecular Evidence for the Link Between Colorectal Cancer and Obesity

Yang Chen
Division of Medical Informatics, Department of EECS
Case Western Reserve University
Cleveland, Ohio, USA
yxc233@case.edu

Li Li, MD, PhD
Departments of Family Medicine and Community Health, Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA
li.li@uhhospitals.org

Rong Xu, PhD
Division of Medical Informatics, Case Western Reserve University
Cleveland, Ohio, USA
rxx@case.edu

***Abstract*— Epidemiological studies suggested that obesity increases the risk of colorectal cancer (CRC). The genetic connection between CRC and obesity is multifactorial and inconclusive. In this study, we hypothesize that the study of shared comorbid diseases between CRC and obesity can offer unique insights into common genetic basis of these two diseases. We constructed a comorbidity network based on mining health data for millions of patients. We developed a novel approach and extracted the diseases that play critical roles in connecting obesity and CRC in the comorbidity network. Our approach was able to prioritize metabolic syndrome and diabetes, which are known to be associated with obesity and CRC through insulin resistance pathways. Interestingly, we found that osteoporosis was highly associated with the connection between obesity and CRC. Through gene expression meta-analysis, we identified novel genes shared among CRC, obesity and osteoporosis. Literature evidences support that these genes may contribute in explaining the genetic overlaps between obesity and CRC.**

Keywords—comorbidity network; colorectal cancer; obesity; osteoporosis; association rule mining; gene expression

I. INTRODUCTION

Comorbidity studies often detect unexpected disease links [1] and offer novel insights into the genetic mechanisms of diseases [2, 3]. A number of epidemiological studies suggest that obesity increases the risk of colorectal cancer (CRC) [4-6]. Based on these evidences of co-occurrence, many genetic factors have been proposed to explain the role of obesity in the development of CRC. For example, both animal and human studies have demonstrated that the increased release of insulin and reduced insulin signaling play roles in obesity and colorectal carcinogenesis [7-9]. Experiments also show that obesity leads to altered level of adipocytokines, such as Adiponectin [10-12] and leptin [13, 14], which may either prevent or foster carcinogenesis.

The mechanism for the association between obesity and CRC is multifactorial and inconclusive [6, 15, 16]. Shared comorbidities between obesity and CRC can provide unique insights into the common genetic basis for the two diseases. For example, type 2 diabetes is highly correlated with obesity and was identified as a risk factor for CRC [17]. A few studies then discovered that genetic factors of insulin resistance, which occur in type 2 diabetes, contribute in explaining the role of obesity in CRC [18]. However, both obesity and CRC are heterogeneous conditions. Over 40% of the obese population is not characterized by the presence of insulin resistance [19]. We hypothesize that systems approaches to studying the diseases that are phenotypically-significant to both CRC and obesity may offer new insights into the common molecular mechanisms between the two interconnected diseases.

Systematic comorbidity studies have been conducted previously, but mostly focused on pairwise comorbidities and their genetic overlaps. Rhetsky et al. developed a statistical model to estimate the co-occurrence relationship for each pair of 160 diseases [20], and demonstrated that comorbidities are genetically linked. Park et al. [21] and Hidalgo et al. [22] detected the comorbidities pairs from the Medicare claims (which only contain senior patients ages 65 or older) with statistical measures. Roque et al. mined pairwise disease correlations using similar measures from medical records of a psychiatric hospital [23]. Recently, we extracted comorbidity patterns from a publically accessible database, which contains disease records for millions of patients at all ages, using an association rule mining approach [24, 25].

In this study, we constructed a disease comorbidity network based on our previous work. We developed a novel approach to detect diseases that have strong connections with both obesity and CRC in the comorbidity network. Specifically, we extracted the local network consisting of all the paths between obesity and CRC, and prioritized the nodes (diseases) that play critical roles in maintaining the connection between the two diseases (Fig.1). Substantial literature evidences can support that the top ranked diseases have associations with both obesity and CRC. We investigated the gene expression profiles of a prioritized comorbid disease to facilitate detecting novel genetic basis underlying the link between obesity and CRC. Our approach is generalizable to study the genetic basis for other disease associations.

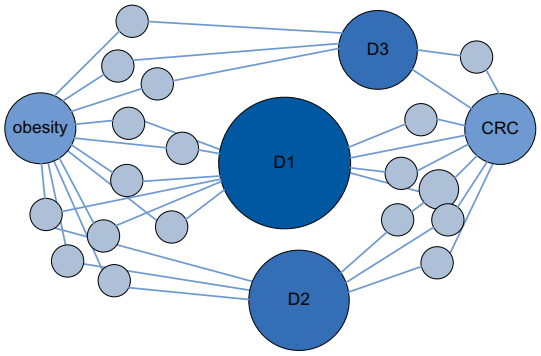


Fig.1. Approach to detect the diseases that have strong connections with both obesity and CRC in the comorbidity network. Nodes D1, D2 and D3 were prioritized because they play important roles in maintaining the network structure and the connection

II. Materials and Methods

Fig.2 shows the three steps of our approach. We first mined disease comorbidity relationships from large amounts of patient records and constructed a disease comorbidity network. We then extracted the local comorbidity cluster for obesity and CRC and prioritize the candidate comorbidity that plays a critical role in connecting the two diseases. Finally we conducted gene expression meta-analysis to identify common genes shared by obesity, CRC and the prioritized comorbidity.

A. Construct Disease Comorbidity Network

We mined disease comorbidity relationships from the FDA adverse event reporting system. The database contains records (2004-2013) of 3,354,043 patients (male and female at all age levels) and 10,112 disorders. Our previous studies [24, 25] have demonstrated that this database is useful in mining comorbidity patterns among diverse patient populations.

We applied the association rule mining approach to detect disease comorbidity relationships from the patient-disease pairs. Association rule mining can flexibly detect strong co-occurrence relationships among sets of diseases, and alleviates the intrinsic bias of traditional comorbidity measures (such as relative risk and φ-correlation) towards rare diseases [24, 25].

We constructed an undirected and unweighted comorbidity network based on the result of association rule mining, which is a list of patterns between two sets of diseases, represented in the form x→y. We collected all diseases in the set x and y in each pattern, assuming they have comorbidity relationships with each other, and established an edge between each pair of diseases in xUy to construct the comorbidity network [24].

B. Prioritize the Diseases That Have Strong Associations with Both Obesity and CRC

We extracted the local network consisting of the paths from obesity to CRC in the disease comorbidity network. The local network thus includes the nodes that may represent different aspects of the relationship between obesity and CRC. We implemented breath first search to enumerate the paths, and limited the paths within four steps.

Then we ranked the nodes in the local network, except obesity and CRC, based on how important they are in maintaining the local network structure and the connection between obesity and CRC. We used the degree and betweenness centrality to characterize the importance of each node in the flowing of the network. The degree of a node becomes higher if more paths between obesity and CRC pass through this node. The betweeness evaluates the number of times that the node acts as the bridge along the shortest paths. Removing the nodes with highest degree or betweenness can easily break down the connection between obesity and CRC.

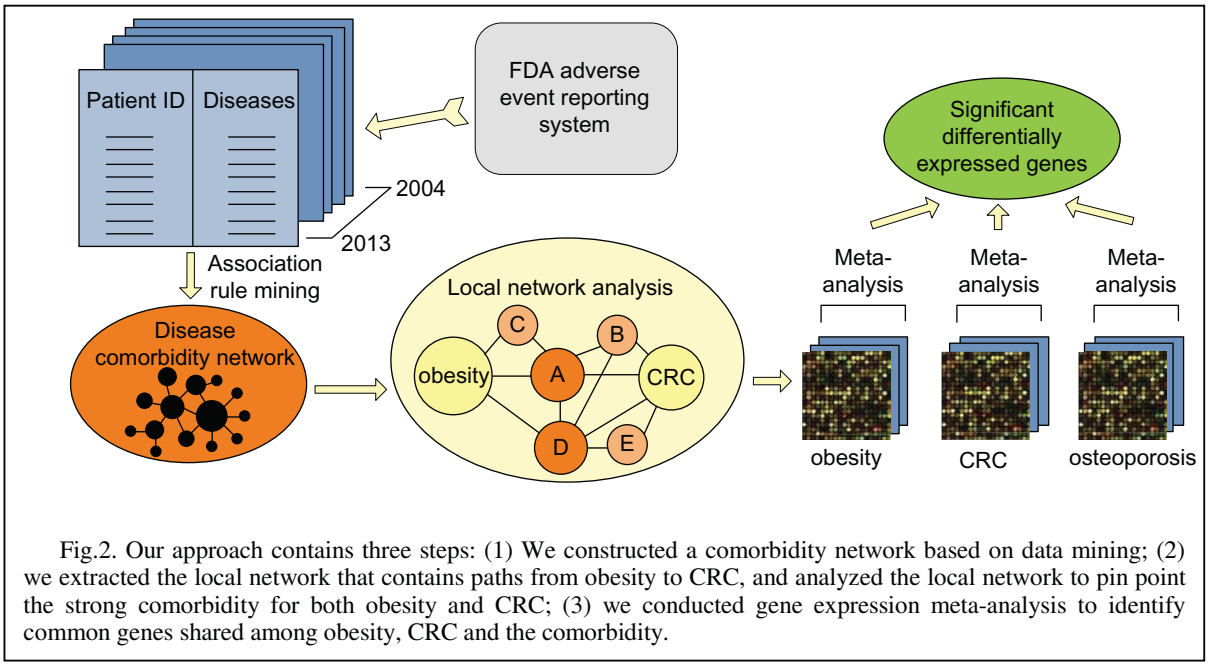


Fig.2. Our approach contains three steps: (1) We constructed a comorbidity network based on data mining; (2) we extracted the local network that contains paths from obesity to CRC, and analyzed the local network to pin point the strong comorbidity for both obesity and CRC; (3) we conducted gene expression meta-analysis to identify common genes shared among obesity, CRC and the comorbidity.

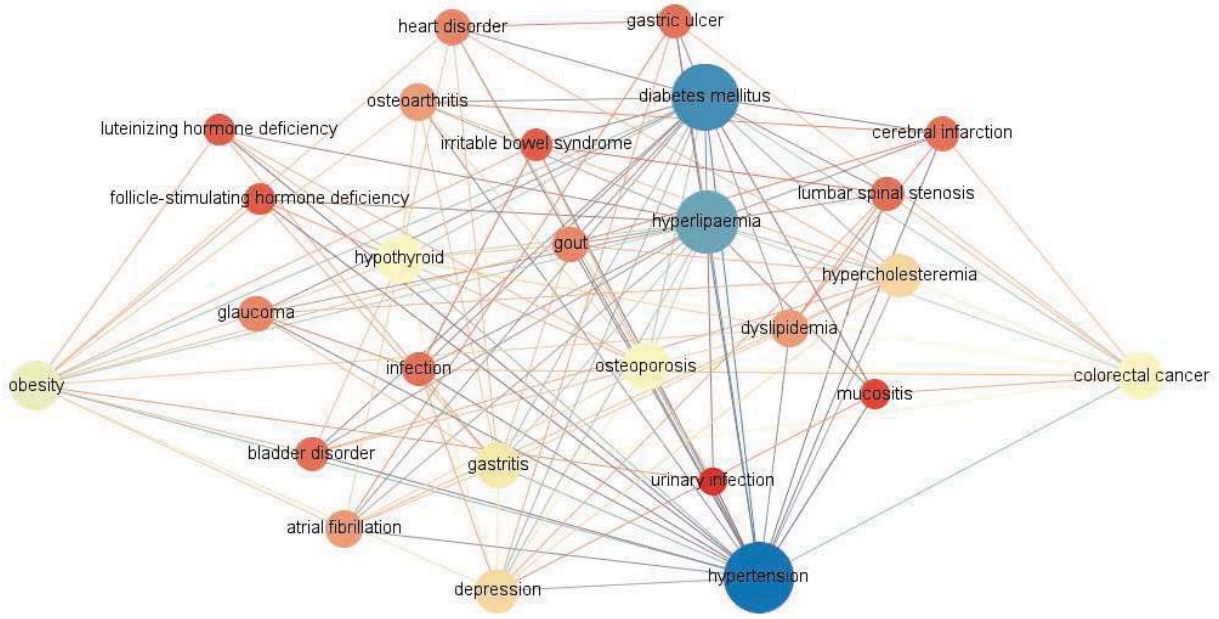


Fig.3. The local network that contains all paths from obesity to colorectal cancer in the comorbidity network.

We investigated the top ranked diseases based on both ranking methods, and used the unexpected ones to guide the detection of genetic associations between obesity and CRC.

C. *Identify Gene Overlaps Through Gene Expression Meta-analysis*

We chose a top ranked disease on the path between obesity and CRC, and then conducted gene expression meta-analysis for the prioritized disease, obesity and CRC, respectively, to detect new genetic explanations for the relationship between obesity and CRC. Gene expression normalized data (SOFT files) were downloaded from NCBI GEO omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) using the R package GEOquery [28]. Then, we performed microarray meta-analyses for each disease independently using the R package MetaDE [29]. MetaDE implements meta-analysis methods for differential expression analysis, and we used the Fisher's method. Significant differentially expressed genes (DEGs) were selected as those displaying a FDR corrected p-value <0.05. Last, we extracted the common significant genes for the three diseases.

III. Results

A. *Local Disease Comorbidity Network Models the Connection Between Obesity and CRC*

We extracted 7006 comorbidity association rules with the confidence larger than 50% from the patient records across ten years. The comorbidity network based on these rules contains 771 nodes and 15,667 edges. Fig.3 shows the local network consisting of all the 119 paths (no longer than four steps) from obesity to CRC. A total of 24 nodes in the local network are the candidate diseases, which have associations with both obesity and CRC, and may indicate different aspects of the relationship between the two diseases.

B. *Osteoporosis Shows High Comorbidity Associations with Both CRC and Obesity*

Table 1 shows the top five nodes sorted by degree and betweenness in the local network. In either way of ranking, hypertension, diabetes and hyperlipaemia were in top three and closely related with both obesity and CRC. Substantial literature evidences support that the metabolic syndrome components, hypertension and hyperlipaemia, as well as diabetes have association with obesity and CRC through insulin resistance in substantial literature [6-9, 18]. These three disorders also independently increase the risk of CRC and colorectal adenoma [6, 17, 18]. The top ranked comorbidities demonstrated the validity of our network analysis approach.

Significantly, osteoporosis was ranked highly by both

TABLE I. Top five disease nodes in the local network that contains all paths from obesity to colorectal cancer. The diseases were ranked by degree and betweenness, respectively.

Rank	Ranked by degree		Ranked by betweenness	
	Nodes	*Degree*	*Nodes*	*Betweenness*
1	Hypertension	26	Hypertension	60.2
2	Diabetes mellitus	24	Diabetes mellitus	55.9
3	Hyperlipaemia	22	Hyperlipaemia	35.2
4	Osteoporosis	14	Osteoporosis	12.3
5	Hypothyroid	14	Hypothyroid	9.5

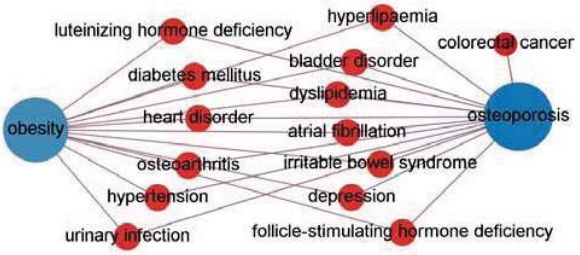


Fig.4. The paths from obesity to colorectal cancer that pass through osteoporosis.

centrality ranking methods. Epidemiological studies suggested an inverse association between bone mineral density and CRC [30], colon cancer among postmenopausal women [31], and colorectal adenoma [32]. On the other hand, patients of obesity and osteoporosis may share common genetic and environmental factors [33]. Different from previous studies, our result shows that osteoporosis is crucial for the association between CRC and obesity. Fig.4 shows the paths of obesity-osteoporosis-CRC. We further investigate the gene expression profiles of osteoporosis patients to gain novel insight of the genetic basis for the link between obesity and CRC.

C. Innovative Genes Shared Among Osteoporosis, Obesity and CRC Were Detected Using Gene Expression Meta-analysis

We downloaded five microarray series (GSE4017, GSE9348, GSE4183, GSE8671, GSE20916) for CRC, three (GSE48964, GSE29718, GSE55205) for obesity and three (GSE7429, GSE2208, GSE7158) for osteoporosis. Through meta-analysis, we obtained 9058 significant differentially expressed genes for CRC, 275 for obesity and 91 for osteoporosis. CRC and obesity shared a total of 192 genes. Among them, we found genes on insulin signaling pathways, such as PDK1, PRKAG2 and PDE3B, and adipocytokines, such as IL6 and IL8.

The three diseases osteoporosis, obesity and CRC shared six genes. Table II lists the genes and literature evidences, which support their relationships with each of the three diseases. Among them, FOS, JUN, and FOSB are oncogenes. FOS and JUN are known on the insulin signaling pathway. FOSB is on the AP1 pathway, which is associated with the proliferation of colon cancer cells [55]. Several studies suggested that overexpression of FOSB increases the responding of high fat reward while decreases energy expenditure and promotes adiposity [40, 56].

Interestingly, we found several genes not involving insulin signaling. Gene PPP1R15A is in the bone morphogenetic

TABLE II. COMMON GENES SHARED BY OBESITY, CRC AND OSTEOPOROSIS, AND PLAUSIBLE EVIDENCE SUPPORTING THEIR RELATIONSHIPS WITH THE THREE DISEASES.

GENES	OBESITY	CRC	OSTEOPOROSIS
PPP1R15A*	In the bone morphogenetic protein (BMP) signaling pathway, which regulates appetite [34]	Mutations in the BMP pathway are related with colorectal carcinogenesis [35]	In the bone morphogenetic protein signaling pathway, which are associated with bone-related diseases, such as osteoporosis [36]
FOS	diet-induced obesity is accompanied by alteration of FOS expression [37]	Proto-oncogene, in the KEGG pathway of colorectal cancer [38]	Mice lacking c-fos develop severe osteopetrosis [39]
FOSB	positive association between maternal obesity [40]	Oncogene, regulators of cell proliferation, has a debatable impact on CRC patient survival [41]	Overexpression of FosB increases bone formation [42]
HADHA*	Associated with multiple fatty acid metabolism pathways [43]	Unknown. Associated with breast cancer [44]	Unknown.
JUN	The c-Jun NH2-terminal Kinase Promotes Insulin Resistance [45]	Proto-oncogene, in the KEGG pathway of colorectal cancer [38]	Associated with osteogenesis [46, 47]
NRIP1*	Down-regulated in obese subjects, may suggest a compensatory mechanism to favor energy expenditure and reduce fat accumulation in obesity states [48]	Unknown. Involved in regulation of E2F1, an oncogene [49]	Modulates transcriptional activity of the estrogen receptor. Interact with ESR1 and ESR2 in osteoporosis [50]

* novel genes not involving insulin resistance pathways

protein signaling (BMP) pathway and its superfamily, the TGF beta signaling pathway. The mutation of BMP pathway has been found in patients with juvenile polyposis, which is rare syndrome with an increased risk for developing CRC [51, 52]. Mutations in TGF beta signaling also have been found susceptibility to CRC through genome-wide association studies [53]. A recent mouse experiment also showed that the BMP pathway regulates brown adipogenesis, energy expenditure and appetite, thus is highly associated with diet-induced obesity [54]. These evidences support our result. Further investigation is required to confirm and elucidate the role of the BMP pathway in the connection between obesity and CRC.

Gene NRIP1 regulates the estrogen receptor. Its interaction with sex hormone receptors plays a role in both obesity [48] and osteoporosis [50]. Its relationship with CRC is unclear yet, but studies suggested that estrogen may have protective effect on CRC [57]. Gene HADHA is on multiple pathways of fatty acid metabolism. But its role in CRC and osteoporosis is unknown yet.

To identify the common genes among obesity, CRC and osteoporosis, we currently analyzed the gene expression data, which can be noisy. While we found literature evidences to support the detected genes and their relationships with both obesity and CRC, these candidate genes need further investigations, for example, through mouse model experiments.

IV. Discussions and Conclusions

The genetic connection between CRC and obesity is multifactorial and inconclusive. In this study, we developed a comorbidity network analysis approach, which suggested that osteoporosis is important for the connection between obesity and CRC. We identified common genes among obesity, CRC and osteoporosis, and found these genes are associated with the regulation of sex hormone receptors and growth factors inducing bone formation. These genes are candidates in explaining the genetic overlaps between obesity and CRC.

Our comorbidity network may be not inclusive and biased toward the diseases whose drugs have high toxicity. The FDA adverse event reporting system collects data from medical product manufacturers, health professionals, and the public. The diseases without drug treatments are not included in the data, and the disease comorbidity relationships were often under-estimated in practice based on these data. In this study, we developed a network analysis approach to compensate the bias of the comorbidity data. In the future, including more complete patient disease data may facilitate the detection of new interesting comorbidities other than osteoporosis for obesity and CRC.

In addition, we currently detect comorbidities based on disease co-occurrence. The co-occurrence patterns may indicate the increase of the risk between two diseases in a mutual way. Incorporating more comprehensive patient-level data, such as time series data, may help refine the disease relationships and control confounding factors.

Acknowledgment

Our research was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award number DP2HD084068, the training grant in computational genomic epidemiology of cancer (CoGEC) (R25 CA094186-06)

References

[1] M. Oti, M. A. Huynen and H. G. Brunner. Phenome connections, Trends Genet, 24(3), 103–106.

[2] Blair, D. R., Lyttle, C. S., Mortensen, J. M., et al. (2013). A nondegenerate code of deleterious variants in mendelian Loci contributes to complex disease risk. *Cell*, **155**(1), 70–80.

[3] Avery, C. L., He, Q., North, K. E., et al. (2011). A phenomics-based strategy identifies loci on APOC1, BRAP, and PLCG1 associated with metabolic syndrome phenotype domains. *PLoS genetics*, **7**(10), e1002322.

[4] Calle, E. E., Rodriguez, C., Walker-Thurmond, K., and Thun, M. J. (2003). Overweight, obesity, and mortality from cancer in a prospectively studied cohort of US adults. New England Journal of Medicine, 348(17), 1625-1638.

[5] Bardou, M., Barkun, A. N., and Martel, M. (2013). Obesity and colorectal cancer. Gut, 62(6), 933-947.

[6] Khaodhiar, L., McCowen, K. C., and Blackburn, G. L. (1999). Obesity and its comorbid conditions. Clinical cornerstone, 2(3), 17-31.

[7] Pollak, M. (2008). Insulin and insulin-like growth factor signalling in neoplasia. Nature Reviews Cancer, 8(12), 915-928.

[8] LeRoith, D., and Roberts Jr, C. T. (2003). The insulin-like growth factor system and cancer. Cancer letters, 195(2), 127-137.

[9] Renehan, A. G., Zwahlen, M., Minder, C., O'Dwyer, S. T., Shalet, S. M., and Egger, M. (2004). Insulin-like growth factor (IGF)-I, IGF binding protein-3, and cancer risk: systematic review and meta-regression analysis. The Lancet, 363(9418), 1346-1353.

[10] Dalamaga, M., Diakopoulos, K. N., and Mantzoros, C. S. (2012). The role of adiponectin in cancer: a review of current evidence. Endocrine reviews, 33(4), 547-594.

[11] An, W., Bai, Y., Deng, S. X., Gao, J., Ben, Q. W., Cai, Q. C., ... and Li, Z. S. (2012). Adiponectin levels in patients with colorectal cancer and adenoma: a meta-analysis. European Journal of Cancer Prevention, 21(2), 126-133.

[12] Wei, E. K., Giovannucci, E., Fuchs, C. S., Willett, W. C., and Mantzoros, C. S. (2005). Low plasma adiponectin levels and risk of colorectal cancer in men: a prospective study. Journal of the National Cancer Institute, 97(22), 1688-1694.

[13] Pärstattin, R. P., Söderberg, S., Biessy, C., Ardnor, B., Kaaks, G. H. R., and Olsson, T. (2003). Plasma leptin and colorectal cancer risk: a prospective study in Northern Sweden. Oncology reports, 10, 2015-2021.

[14] Tamakoshi, K., Toyoshima, H., Wakai, K., Kojima, M., Suzuki, K., Watanabe, Y., ... and Tamakoshi, A. (2005). Leptin is associated with an increased female colorectal cancer risk: a nested case-control study in Japan. Oncology, 68(4-6), 454-461.

[15] Zhu, Y., Michelle Luo, T., Jobin, C., and Young, H. A. (2011). Gut microbiota and probiotics in colon tumorigenesis. Cancer letters, 309(2), 119-127.

[16] Danese, E., Montagnana, M., Minicozzi, A. M., Bonafini, S., Ruzzenente, O., Gelati, M., ... and Guidi, G. C. (2012). The role of resistin in colorectal cancer. Clinica Chimica Acta, 413(7), 760-764.

[17] Berster, J. M., and Göke, B. (2008). Type 2 diabetes mellitus as risk factor for colorectal cancer. Archives of physiology and biochemistry, 114(1), 84-98.

[18] Komninou, D., Ayonote, A., Richie, J. P., and Rigas, B. (2003). Insulin resistance and its contribution to colon carcinogenesis. Experimental Biology and Medicine, 228(4), 396-405.

[19] Karelis, A. D. (2008). Metabolically healthy but obese individuals. Lancet, 372(9646), 1281-1283.

[20] Rzhetsky A, Wajngurt D, Park N, and Zheng T. Probing genetic overlap among complex human phenotypes. Proc Natl Acad Sci USA. 2007;104:11694–9.

[21] Park J, Lee DS, Christakis NA, and Barabasi AL. The impact of cellular networks on disease comorbidity. Mol Syst Biol. 2009;5:262.

[22] Hidalgo CA, Blumm N, Barabási AL, Christakis NA. A dynamic network approach for the study of human phenotypes. PLoS Comput Biol. 2009;5: e1000353.

[23] Roque FS, Jensen PB, Schmock H, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol. 2011;7(8):e1002141.

[24] Chen, Y., and Xu, R. (2014). Network Analysis of Human Disease Comorbidity Patterns Based on Large-Scale Data Mining. In Bioinformatics Research and Applications (pp. 243-254). Springer International Publishing.

[25] Chen Y, and Xu R (2014) Mining cancer-specific disease comorbidities from a large observational database, Cancer Informatics (in press).

[26] Capobianco, E., and Lio, P. (2013). Comorbidity: a multidimensional approach. Trends in molecular medicine, 19(9), 515-521.

[27] Cramer, A. O., Waldorp, L. J., van der Maas, H. L., and Borsboom, D. (2010). Comorbidity: a network perspective. Behavioral and Brain Sciences, 33(2-3), 137-150.

[28] Davis S and Meltzer P (2007). "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor." Bioinformatics, 14, pp. 1846–1847.

[29] Wang, X., Kang, D. D., Shen, K., Song, C., Lu, S., Chang, L. C., ... and Tseng, G. C. (2012). An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. Bioinformatics, 28(19), 2534-2536.

[30] Nelson RL, Turyk M, Kim J, Persky V. Bone mineral density and the subsequent risk of cancer in the NHANES I follow-up cohort. BMC Cancer. 2002;2:22–31.

[31] Ganry O, Lapotre-Ledoux B, Fardellone P, Dubreuil A. Bone mass density, subsequent risk of colon cancer and survival in postmenopausal women. Eur J Epidemiol. 2008;23:467–73.

[32] Nock, N. L., Patrick‐Melin, A., Cook, M., Thompson, C., Kirwan, J. P., and Li, L. (2011). Higher bone mineral density is associated with a decreased risk of colorectal adenomas. International Journal of Cancer, 129(4), 956-964.

[33] Zhao, L. J., Liu, Y. J., Liu, P. Y., Hamilton, J., Recker, R. R., and Deng, H. W. (2007). Relationship of obesity with osteoporosis. The Journal of Clinical Endocrinology & Metabolism, 92(5), 1640-1646.

[34] Townsend, K. L., Suzuki, R., Huang, T. L., Jing, E., Schulz, T. J., Lee, K., ... & Tseng, Y. H. (2012). Bone morphogenetic protein 7 (BMP7) reverses obesity and regulates appetite through a central mTOR pathway. The FASEB Journal,26(5), 2187-2196.

[35] Hardwick, J. C., Kodach, L. L., Offerhaus, G. J., & Van den Brink, G. R. (2008). Bone morphogenetic protein signalling in colorectal cancer. Nature Reviews Cancer, 8(10), 806-812.

[36] Chen, G., Deng, C., & Li, Y. P. (2012). TGF-β and BMP signaling in osteoblast differentiation and bone formation. International journal of biological sciences,8(2), 272.

[37] Parker, J. A., McCullough, K. A., Field, B. C. T., Minnion, J. S., Martin, N. M., Ghatei, M. A., & Bloom, S. R. (2013). Glucagon and GLP-1 inhibit food intake and increase c-fos expression in similar appetite regulating centres in the brainstem and amygdala. International Journal of Obesity, 37(10), 1391-1398.

[38] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

[39] Okada, S., Wang, Z. Q., Grigoriadis, A. E., Wagner, E. F., & von Rüden, T. (1994). Mice lacking c-fos have normal hematopoietic stem cells but exhibit altered B-cell differentiation due to an impaired bone marrow environment.Molecular and cellular biology, 14(1), 382-390.

[40] Thakali, K. M., Saben, J., Faske, J. B., Lindsey, F., Gomez-Acevedo, H., Lowery Jr, C. L., ... & Shankar, K. (2014). Maternal Pre-Gravid Obesity Changes Gene Expression Profiles Towards Greater Inflammation and Reduced Insulin Sensitivity in Umbilical Cord. Pediatric research.

[41] Pfannschmidt, J., Bade, S., Hoheisel, J., Muley, T., Dienemann, H., & Herpel, E. (2009). Identification of immunohistochemical prognostic markers for survival after resection of pulmonary metastases from colorectal carcinoma. The Thoracic and cardiovascular surgeon, 57(7), 403-408.

[42] Sabatakos, G., Sims, N. A., Chen, J., Aoki, K., Kelz, M. B., Amling, M., ... & Baron, R. (2000). Overexpression of ΔFosB transcription factor (s) increases bone formation and inhibits adipogenesis. Nature medicine, 6(9), 985-990.

[43] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0.Bioinformatics, 27(12), 1739-1740.

[44] Mamtani, M., & Kulkarni, H. (2012). Association of HADHA expression with the risk of breast cancer: targeted subset analysis and meta-analysis of microarray data. BMC research notes, 5(1), 25.

[45] Aguirre, V., Uchida, T., Yenush, L., Davis, R., & White, M. F. (2000). The c-Jun NH2-terminal kinase promotes insulin resistance during association with insulin receptor substrate-1 and phosphorylation of Ser307. Journal of Biological Chemistry, 275(12), 9047-9054.

[46] Lewinson, D., Rachmiel, A., Rihani-Bisharat, S., Kraiem, Z., Schenzer, P., Korem, S., & Rabinovich, Y. (2003). Stimulation of Fos-and Jun-related genes during distraction osteogenesis. Journal of Histochemistry & Cytochemistry,51(9), 1161-1168.

[47] Krzeszinski, J. Y., Wei, W., Huynh, H., Jin, Z., Wang, X., Chang, T. C., ... & Wan, Y. (2014). miR-34a blocks osteoporosis and bone metastasis by inhibiting osteoclastogenesis and Tgif2. Nature, 512(7515), 431-435.

[48] Catalán, V., Gómez-Ambrosi, J., Lizanzu, A., Rodríguez, A., Silva, C., Rotellar, F., ... & Frühbeck, G. (2009). RIP140 Gene and protein expression levels are downregulated in visceral adipose tissue in human morbid obesity. Obesity surgery, 19(6), 771-776.

[49] Docquier, A., Harmand, P. O., Fritsch, S., Chanrion, M., Darbon, J. M., & Cavaillès, V. (2010). The transcriptional coregulator RIP140 represses E2F1 activity and discriminates breast cancer subtypes. Clinical Cancer Research,16(11), 2959-2970.

[50] Morón, F. J., Mendoza, N., Vázquez, F., Molero, E., Quereda, F., Salinas, A., ... & Ruiz, A. (2006). Multilocus analysis of estrogen-related genes in Spanish postmenopausal women suggests an interactive role of< i> ESR1</i>,< i> ESR2</i> and< i> NRIP1</i> genes in the pathogenesis of osteoporosis. Bone,39(1), 213-221.

[51] Howe, J. R., Bair, J. L., Sayed, M. G., Anderson, M. E., Mitros, F. A., Petersen, G. M., ... & Vogelstein, B. (2001). Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. Nature genetics, 28(2), 184-187.

[52] Brosens, L. A., van Hattem, A., Hylind, L. M., Iacobuzio-Donahue, C., Romans, K. E., Axilbund, J., ... & Giardiello, F. M. (2007). Risk of colorectal cancer in juvenile polyposis. Gut, 56(7), 965-967.

[53] Bellam, N., & Pasche, B. (2010). TGF-β signaling alterations and colon cancer. In Cancer Genetics (pp. 85-103). Springer US.

[54] Townsend, K. L., Suzuki, R., Huang, T. L., Jing, E., Schulz, T. J., Lee, K., ... & Tseng, Y. H. (2012). Bone morphogenetic protein 7 (BMP7) reverses obesity and regulates appetite through a central mTOR pathway. The FASEB Journal, 26(5), 2187-2196.

[55] Ashida, R., Tominaga, K., Sasaki, E., Watanabe, T., Fujiwara, Y., Oshitani, N., ... & Arakawa, T. (2005). AP-1 and colorectal cancer. Inflammopharmacology, 13(1-3), 113-125.

[56] Vialou, V., Cui, H., Perello, M., Mahgoub, M., Yu, H. G., Rush, A. J., ... & Lutter, M. (2011). A role for ΔFosB in calorie restriction-induced metabolic changes. Biological psychiatry, 70(2), 204-207.

[57] Barzi, A., Lenz, A. M., Labonte, M. J., & Lenz, H. J. (2013). Molecular pathways: estrogen pathway in colorectal cancer. Clinical Cancer Research, 19(21), 5842-5848.

Development of Bioinformatics Pipeline for Analyzing Clinical Pediatric NGS Data

Erin L. Crowgey, MS[1], Anders Kolb, MD[2], and Cathy H. Wu, PhD[1]

[1]Center for Bioinformatics & Computational Biology, University of Delaware, Newark, DE;
[2]Nemours Alfred I. DuPont Hospital for Children, Wilmington, DE

Abstract
Using an Illumina exome sequencing dataset generated from pediatric Acute Myeloid Leukemia patients (AML; type FLT3/ITD+) a comprehensive bioinformatics pipeline was developed to aid in a better clinical understanding of the genetic data associated with the clinical phenotype. The pipeline starts with raw next generation sequencing reads and using both publicly available resources and custom scripts, analyzes the genomic data for variants associated with pediatric AML. By incorporating functional information such as Gene Ontology annotation and protein-protein interactions, the methodology prioritizes genomic variants and returns disease specific results and knowledge maps. Furthermore, it compares the somatic mutations at diagnosis with the somatic mutations at relapse and outputs variants and functional annotations that are specific for the relapse state.

Introduction
Acute myeloid leukemia (AML) is a complex disease characterized by dysregulation of signal transduction pathways in hematopoietic progenitors that ultimately results in the increase of proliferation and survival of leukemic cells [(1)]. AML is considered a disease of the genome as many genetic alterations are required for onset. Genomic variants for AML are often described as either Type I mutations, which alter cell proliferation, or Type II mutations, which alter cell survival pathways [(2)].

Pediatric AML is a rare disease with only ~500 children a year diagnosed (stjude.org) and prognosis has improved over the decades[(3)]. However, relapse is a major concern and accounts for more than half of the deaths in pediatric leukemia cases [(1)(3)]. Common mutations associated with AML are found in several genes including FLT3, NPM1, CEBPA, RAS, c-KIT, and WT1. Furthermore, co-occurring mutations such as an internal tandem duplication (ITD) in the FLT3 gene accompanied by mutations in WT1, have been associated with poor outcome (4). The FLT3/ITD is an in-frame insertion in exon 14 or 15 that changes the amino acid sequence in the juxtamembrane domain, leading to ligand-independent FLT3 activation (5). In the clinical setting FLT3 / ITD is detected through a PCR based assay, and additional testing is required to further analyze the sample for other potential genomic mutations.

Recent advancements in DNA sequencing technology have aided in our ability to detect numerous genetic alterations from a single genomic sample. These advancements can aid in personalized medicine by revealing the genomic architecture of a specific patient. However, applying NGS in the medical field requires knowledgeable personnel and significant computer infrastructure and algorithms specific for handling the large datasets. This study retrospectively analyzes FLT3/ITD positive samples, diagnosis, remission, and relapse, with the goal of developing a bioinformatics pipeline capable of detecting the FLT3/ITD, along with other genetic alterations, which collectively can aid in a better understanding of biological processes dysregulated in the relapse state of pediatric AML. The goal of the bioinformatics pipeline is to provide an enhanced output that allows a clinician to better understand the pathways and biological processes affected by the detected genetic alterations.

Starting with raw NGS sequencing reads, bioinformatics pipelines were created for analyzing exon-captured Illumina data. The pipeline combines publicly available algorithms and custom scripts to detect and prioritize genomic variants. Six FLT3/ITD positive pediatric AML samples, with varying FLT3/ITD allelic ratios, were analyzed using the developed methodologies. A thorough analysis between the diagnosis and relapse sample was conducted for each patient, revealing several relapse specific mutations. Our pipeline detected different types of genetic alterations, i.e. large insertions and single nucleotide polymorphisms (SNP), helping to establish NGS as a feasible methodology in the clinical setting. The pipeline is being designed with the flexibility to integrate other genomic detection algorithms in the future, such as copy number variation.

Methods
Illumina paired-end exon-sequencing data generated from bone marrow samples was received from the Children's Oncology Group. The quality of the sequence reads was examined using fastqc (Babraham Institute) and cutadapt

(https://code.google.com/p/cutadapt/) was used to trim low quality bases. The trimmed NGS reads were aligned to the human reference genome (hg19) using bwa-mem (version bwa-0.7.4) (6). Average depth of coverage per exon (vertical) and average exon coverage (horizontal) were calculated using a custom script and Ensembl annotation files. Following the best practices described by the Genome Analysis Tool Kit (GATK) developers (7), alignment files were processed using Picard Tools Version 1.67 (http://picard.sourceforge.net/).

Mutect (Version 1.1.4) and Shimmer (Version 5.8.8) were executed for SNP detection, and to aid in the validation of the pipeline the results were compared to verified variants provided by the Children Oncology Group. Variant call files (VCF version 4.1) were annotated with SnpEff Version 3.3a using the package GRCh37.75 annotation recommended by SnpEff [4]. Using SnpEff annotated transcript ID, variants in the VCFs were mapped to UniProt Accession Numbers and Gene Ontology information using a custom script. Protein-protein interactions for the protein coding genes were determined using the STRING API (9), a database of known and predicted protein interactions derived from: genomic context, high-throughput experiments, co-expression, and previous knowledge.

Pindel algorithm was executed on the alignment files for detection of FLT3/ITD (10). The output files were converted to vcf files using the pindel2vcf script provided with the Pindel package. Only insertions located in exon 14 or 15 in the FLT3 gene were analyzed as potential ITDs. All computational work was performed at the University of Delaware on the BioHen high performance computing cluster.

Results and Discussion

Six FLT3/ITD positive samples, with varying allelic ratios and cytogenetic markers were analyzed with a custom pipeline (Figure 1). The pipeline consisted of publicly available algorithms, such as bwa and GATK, plus custom scripts. A key aspect of the established methodologies is the modularization of algorithms and scripts, which creates an environment that allows for the dynamic integration with up-dated algorithms and databases.

Genomic Detection FLT3/ITD

Detecting large insertions, deletions, and tandem duplications from NGS is a challenging task with only a few high quality algorithms publicly available. Recently, Spencer et al. (11) compared several algorithms for the detection of FLT3/ITD and published that Pindel (10), a pattern growth approach, successfully identified FLT3 / ITD. The Pindel algorithm was incorporated into the pipeline for the detection of an insert in exon 14 or 15 in the FLT3 gene that was consistent with the clinical FLT3/ITD (Table 1). For the 6 patients analyzed, 3 samples per patient, Pindel detected an insert in 5 of the 6 patient's diagnosis sample (83%). The Spencer et al. study reported 100% detection of the FLT3 / ITD in the samples analyzed in their study using a targeted NGS approach (27 genes). For the study presented whole exon-sequencing was used and therefore the coverage in the region of interest was much lower, perhaps decreasing the ability to detect the ITD. Pindel also detected an insert in 4 of the relapse samples and 1 of the remission samples. A benefit to using Pindel is that it provides a better resolution of the genomic abnormality by providing a genomic position, sequence, and length of the insert, which are not all available with the PCR electrophoresis assay. Future work for this portion of the pipeline will include an allelic ratio calculation for the FLT3/ITD. This is a difficult task as purity of the cell population sequenced is difficult to determine.

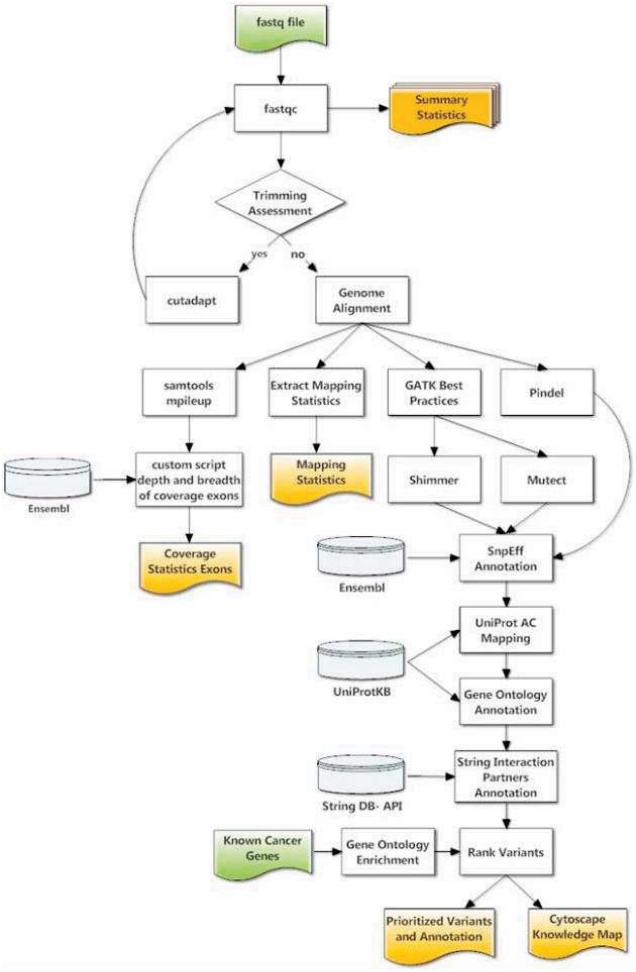


Figure 1. Bioinformatics workflow

Table 1. Summary of Pindel Results

ID	Sample	Position	Sequence	Length
	Diagnosis	-	None Detected	-
Patient 1	Relapse	-	None Detected	-
	Remission	-	None Detected	-
	Diagnosis	28,608,235	TCTTGGAAACTCCCATTTGAGATCATATTCA	31
Patient 2	Relapse	-	None Detected	-
	Remission	-	None Detected	-
	Diagnosis	28,608,249	ATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGCC	40
Patient 3	Relapse	28,608,265	ATATTCTCTGAAATCTCCACGGGGG	25
	Remission	-	None Detected	-
	Diagnosis	28,608,214	CTTACCAAACTCTAAATTTTCTCTTGGAAACTCCCAT	37
Patient 4	Relapse	28,608,214	CTTACCAAACTCTAAATTTTCTCTTGGAAACTCCCAT	37
	Remission	-	None Detected	-
	Diagnosis	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
Patient 5	Relapse	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
	Remission	28,608,223	CTCTAAATTTTCTCTTGGAAACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTA	76
	Diagnosis	28,608,243	ACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCAC	73
Patient 6	Relapse	28,608,243	ACTCCCATTTGAGATCATATTCATATTCTCTGAAATCAACGTAGAAGTACTCATTATCTGAGGAGCCGGTCAC	73
	Remission	-	None Detected	-

Genomic Detection Somatic SNPs and InDels

The pipeline is composed of 3 genomic variant detection algorithms, Pindel, Mutect, and Shimmer, that collectively report single nucleotide polymorphisms (SNP), small insertions and deletions (InDel), and large InDels. When analyzing cancer samples it is important to distinguish, and prioritize, somatic SNPs versus germline SNPs. To aid with validating the pipeline, the somatic SNPs detected were fist compared to the list of verified variants provided by COG (Figure 2). Mutect detected 100% of the verified variants in eight of the twelve samples (diagnosis and relapse), with an average of 93% detection of verified variants. Shimmer detected 100% of the verified variants in five of the twelve samples, with an average of 78% detection of verified variants.

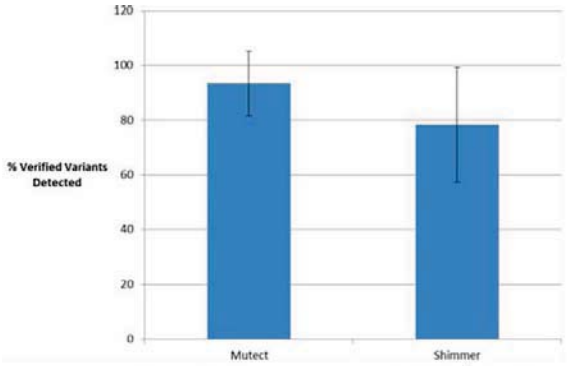


Figure 2. Summary somatic SNP detection

The pipeline also detected high quality somatic SNPS that were not reported to COG. Table 2 summarizes the number of somatic SNPs detected in each sample. For the majority of the patients the relapse samples had more somatic mutations compared with their matched diagnosis sample. Three of the patients had an extremely high number of somatic mutations in their relapse sample, and are undergoing further analysis to determine the potential driver of these mutations.

Table 2. Summary ranked somatic SNPs

ID	Sample	Somatic SNPs	Ranked Variants
Patient 1	diagnosis	201	50
	relapse	16782	9706
Patient 2	diagnosis	160	47
	relapse	7311	1617
Patient 3	diagnosis	177	47
	relapse	111	42
Patient 4	diagnosis	194	71
	relapse	2	50
Patient 5	diagnosis	155	52
	relapse	200	79
Patient 6	diagnosis	153	29
	relapse	7177	4306

Genomic Variant Prioritization

A custom prioritization module was developed to rank the somatic variants, located in protein coding regions of the genome, at the diagnosis state and relapse state using a similar method as described by Hu et al. (12). Five major criteria were used for prioritizing the variants detected: protein-protein interactions, gene ontology, functional consequence, and quality of variant. Using the 27 genes published by Spencer et al., a Gene Ontology (GO) enrichment analysis was done using Bingo, a Cytoscape plug-in. These 27 genes were used because they are cited as genes with known genetic alterations associated with pediatric AML. GO terms that had a significant p-value (<0.05) were extracted, and variants located in a gene annotated with one of the enriched GO terms, were given a positive score. Protein-protein interactions were scored with a similar strategy, with positive scores given to variants located in a gene whose product has a protein-protein interaction with

a protein known to be associated with pediatric AML. The goal is to use characteristics of known pediatric AML cancer genes to identify new genes of interest.

Cytoscape Knowledge Maps

A comparison between diagnosis and relapse samples was performed to better understand shifts in the biological processes influenced by genetic alterations between the two time points. A special feature of the pipeline is the automatic generation of a knowledge map consisting of protein-protein interactions and GO terms associations for the genes of interest that can be easily displayed in Cytoscape.

The Cytoscape map displays genes with a highly ranked mutation as red nodes connected to their GO term (green nodes) and other interacting proteins or proteins with a genetic alteration that does not cause a change in amino acid sequence (blue nodes). Figure 2 highlights an example map generated for a Patient's relapse state, highlighting mutations specific for the relapse state, except for the FLT3/ITD. The pipeline detected and prioritized variants detected in PAK2, PRAMEF1, PRAMEF13, and RHPN2. There were two variants detected in PAK2 (rs76714248, MAF 0.019 and rs67093638) that are predicted to alter the amino acid sequence of the translated protein. PAK2 is a protein kinase involved in several signaling pathways such as apoptosis and proliferation. Two other genes, PRAMEF1 and PRAMEF13, which mapped to GO negative regulation of apoptosis, were also prioritized for this sample. These genes are also annotated with negative regulation of cell differentiation, negative regulation of retinoic acid receptor signaling, positive regulation of cell proliferation, and negative regulation of transcription. The goal of this type of functional output is to help researchers and clinicians make hypothesis regarding the changes from the diagnosis state to the relapse state. For example, this patient gained several mutations in genes involved in apoptosis, cellular differentiation, and retinoic acid signaling that may alter their susceptibility to treatment.

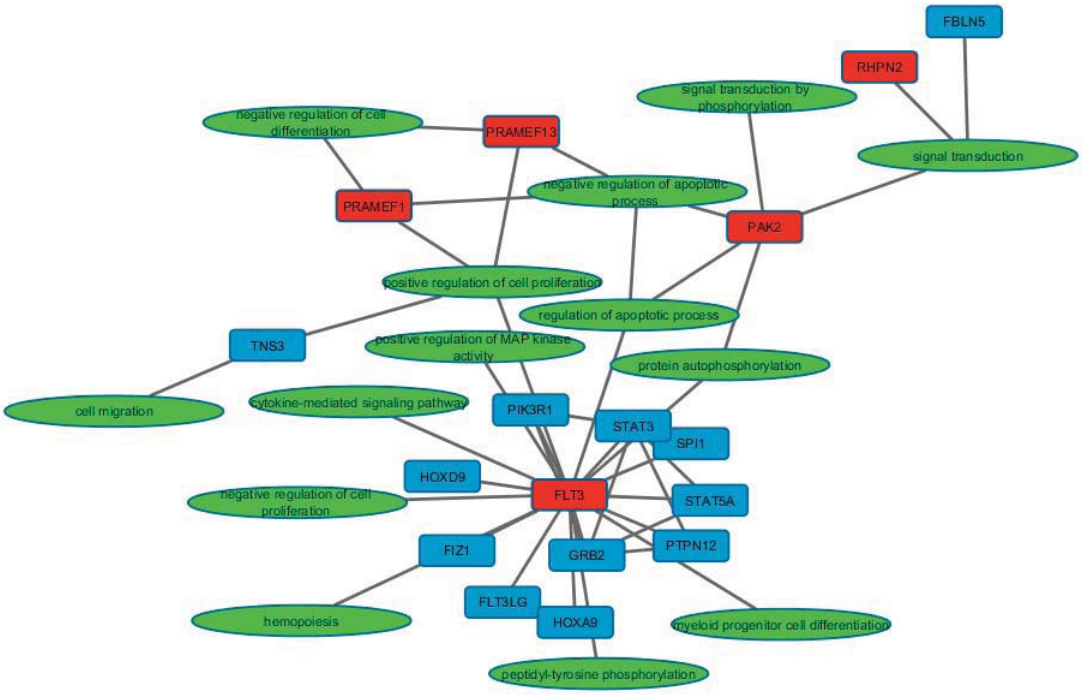


Figure 2. Custom Knowledge Map for Patient 1 Relapse Sample

Cancer is a disease of the genome making it a necessity to be able to analyze multiple types of genetic alterations at once. The application of next generation sequencing has the potential to aid in the diagnosis and treatment of cancer as costs for sequencing decline and the magnitude of data increases. A primary limiting factor to clinical

applications of genomic NGS is downstream bioinformatics analysis. This paper highlights core algorithms required for analyzing clinical NGS samples and reports new algorithms under development for the prioritization and visualization of somatic mutations detected in clinical NGS samples. Currently, the pipeline is available for in-house use only, but in the future it will be made publicly available. Furthermore, as additional samples are analyzed the pipeline will be broaden to rank and distinguish between driver mutations and clinically actionable mutations.

Acknowledgements

This project was partially supported by the Leukemia Research Foundation of Delaware and the Delaware INBRE program, with a grant from the National Institute of General Medical Sciences NIGMS (8 P20 GM103446-13) from the National Institutes of Health. The author appreciates the discussions regarding this project from Chuming Chen, Shawn Polson, and Karen Ross. The author recognizes Karol Miaskiewicz for maintaining the High Performance Cluster (Biohen), and Jennifer Wyffels for proof-reading the manuscript. The author would like to acknowledge Dr. Soheil Meshinchi for his help and guidance with requesting data from the Children's Oncology Group and his input on data analysis.

References

1. Meshinchi S, Arceci RJ. Prognostic factors and risk-based therapy in pediatric acute myeloid leukemia. Oncologist [Internet]. 2007 Mar [cited 2014 Sep 16];12(3):341–55. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17405900
2. Cancer Genomics. Elsevier; 2014.
3. Gamis AS, Alonzo TA, Perentesis JP, Meshinchi S. Children's Oncology Group's 2013 blueprint for research: acute myeloid leukemia. Pediatr Blood Cancer [Internet]. 2013 Jun [cited 2015 Jan 8];60(6):964–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23255301
4. Summers K, Stevens J, Kakkas I, Smith M, Smith LL, Macdougall F, et al. Wilms' tumour 1 mutations are associated with FLT3-ITD and failure of standard induction chemotherapy in patients with normal karyotype AML. Leukemia [Internet]. 2007 Mar [cited 2014 Sep 22];21(3):550–1; author reply 552. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17205055
5. Meshinchi S, Appelbaum FR. Structural and functional alterations of FLT3 in acute myeloid leukemia. Clin Cancer Res [Internet]. 2009 Jul 1 [cited 2014 Sep 22];15(13):4263–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2716016&tool=pmcentrez&rendertype=abstract
6. Li H. Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM. 2013;00(00):1–3.
7. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet [Internet]. 2011 May [cited 2014 Mar 21];43(5):491–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3083463&tool=pmcentrez&rendertype=abstract
8. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) [Internet]. 2012 [cited 2014 Mar 20];6(2):80–92. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3679285&tool=pmcentrez&rendertype=abstract
9. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res [Internet]. 2009 Jan [cited 2014 Mar 28];37(Database issue):D412–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686466&tool=pmcentrez&rendertype=abstract
10. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics [Internet]. 2009 Nov 1 [cited 2014 Jul 14];25(21):2865–71. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2781750&tool=pmcentrez&rendertype=abstract
11. Spencer DH, Abel HJ, Lockwood CM, Payton JE, Szankasi P, Kelley TW, et al. Detection of FLT3 internal tandem duplication in targeted, short-read-length, next-generation sequencing data. J Mol Diagn [Internet]. 2013 Jan [cited 2014 Sep 22];15(1):81–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23159595
12. Hu P, Bader G, Wigle DA, Emili A. Computational prediction of cancer-gene function. Nat Rev Cancer [Internet]. 2007 Jan [cited 2014 Sep 19];7(1):23–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17167517

Conceptualizing a Novel Quasi-Continuous Bayesian Phylogeographic Framework for Spatiotemporal Hypothesis Testing

Daniel Magee[1] and Matthew Scotch PhD, MPH[1]
[1]Arizona State University, Tempe, AZ, USA

Abstract
Continuous phylogeography is a growing approach to studying the spatiotemporal origins of RNA viruses because of its realistic spatial reconstruction advantages over discrete phylogeography. While the generalized linear model has been demonstrated as an effective tool for simultaneously assessing the drivers impacting viral diffusion in discrete phylogeography, there is no similar testing method in the continuous phylogeographic framework. In this paper, we take a step toward bridging that gap by conceptualizing a novel quasi-continuous approach which enables the addition of discrete locations beyond the known sampling locations of the virus. Our model, when fully developed into phylogeographic software, will enable spatiotemporal hypothesis testing of viral diffusion without being strictly limited to observed sampling locations. This model can still assess the impact of local epidemiological variables on virus spread and could provide public health agencies with more realistic estimates of key predictors and locations by utilizing a more continuous landscape.

Introduction
Understanding the spread of disease is one of the most fundamental principles of an effective public health system. This includes the ability to track the spread of disease from location to location while keeping a timescale with respect to the emergence and divergence of viral strains. Phylogeography is an emerging field that has potential for improving public health surveillance. In particular, virus phylogeography considers molecular evolution over geography and has been used to study viral epidemics and pandemics related to influenza[1], West Nile Virus[2], and rabies[3] among others.

The spatial diffusion approach to phylogeography has two main underlying methods: discrete[4] and continuous[3, 5] models. While both are used to estimate virus spread, there are major differences. In discrete phylogeography, ancestral state reconstruction is estimated over observed locations. For example, if viruses are collected in states A, B, C, and D, the model will only consider these when estimating spread. Conversely, in a continuous model spread can be estimated over any sampled or unsampled location, thus alleviating this constraint. In this sense, each ancestral node in a continuous model can have a unique latitude and longitude while the discrete model limits the ancestral nodes to be drawn from the originally defined discrete sites.

In epidemiology, understanding the local variables that play a role in the overall diffusion process of the disease is as vital as the genetic and geographic mechanisms by which they evolve and spread. The rapid development and mutation of viruses, especially those that are RNA-based, makes them especially challenging to fully comprehend. Zoonotic RNA viruses are particularly concerning due to their elevated transmissibility[6, 7] across multiple host species. Aside from the genetic principles alone, there are a variety of variables that have been shown to be risk factors for zoonotic diffusion such as longitude, temperature, and humidity[8], elevation[9], livestock[10, 11] and human population densities[11]. Recent work in the discrete setting has focused on spatiotemporal hypothesis testing for identifying significant predictors of virus spread. For example, work by Lemey et al.[12] demonstrated the use of a generalized linear model (GLM) to identify predictors that contributed to the global spread of H3N2 influenza. The group used a log-linear GLM[13] within a Bayesian framework as seen in (1), below, to reconstruct virus history while simultaneously assessing local predictor variables including passenger flux, population size and density, and latitude.

$$\log(\Lambda_{ij}) = \beta_1\delta_1\log(p_1) + \beta_2\delta_2\log(p_2) + \cdots + \beta_n\delta_n\log(p_n) \quad (1)$$

To evaluate this, data for each predictor variable p was obtained for each discretized location, which can be a state, county, country, or other geographic boundary. A common approach is to utilize the centroid coordinates at each location and gather all predictor data from that point, creating uniformity in the model, and these data are used for hypothesis testing. This group included a binary indicator, δ, to govern the inclusion or exclusion of each given predictor. The specified predictors can be tested for support in the model by a formal Bayes factor test[14] as seen in (2).

$$BF = \frac{p}{1-p} \Big/ \frac{q}{1-q} \quad (2)$$

Here, p is the posterior probability obtained via BEAST[15] for a predictor and q is the prior probability. In this work, the group specified q in which there was a 50% likelihood of no predictor being included in the model. Unsurprisingly, but yet an affirmation of the concept, was the revelation that air travel networks accounted for the most contribution to the diffusion model while other local predictors such as latitude were identified as significant contributors. This demonstrated the usefulness of a GLM while also analyzing the evolutionary history of the virus in a discrete phylogeographic framework and Magee et al.[16] utilized this approach in studying H5N1 in Egypt.

While the GLM has been used successfully in the discrete setting, there are limitations to its implementation for continuous phylogeography. In discrete Bayesian phylogeography, each branch in a phylogeny is suggested to be an independent continuous-time Markov Chain (CTMC)[4] which emit discrete outcomes over a function of continuous time. For K discrete sites in the phylogeny there exists an infinitesimal rate matrix to characterize the CTMC having the distinct property of being stochastic upon exponentiation of the matrix. An eigen decomposition of this matrix can yield the transition probabilities of the rates in finite-time. Continuous phylogeography does not have a defined number of sites, so the K x K matrix does not exist and the GLM principles fail. Instead there is a rate scalar ϕ_b for each branch b of the phylogeny, and the overall precision matrix P is scaled to this ϕ_b. The precision matrix P has two parameters for bivariate diffusion, p_1 and p_2, and also a variable r. Here p_1 and p_2, represent the precision in each spatial dimension, respectively, while r is the correlation coefficient between them. This forms a relaxed random walk model which overcomes a restrictive Brownian diffusion process where each branch in a phylogeny has the same evolutionary rate[3].

A Brownian diffusion model in phylogeography requires all branches of the tree to evolve at the same rate. This assumption is unrealistic and constraining in terms of evolution and needed to be outfitted to better demonstrate evolutionary principles. To overcome this limitation Lemey et al.[3] introduced a bivariate Brownian random walk into the Bayesian framework that accompanies widely used phylogeographic models. This relaxed random walk enabled the individual branches of the phylogeographic trees to have their own evolutionary rate to more accurately portray the underlying evolutionary principles. Furthermore, this avoided an overparameterization issue in the eigen decomposition that comes with having too many sparse discrete locations while allowing additional geographic locations that the observed discrete states.

In a GLM, the number of parameters is dependent upon the number of predictors rather than the number of discrete locations. This avoids the overparameterization issue exhibited with a large set of sparse discrete locations. The roadblock toward incorporating a GLM in the continuous model is the lack of parameters as rates between locations when dealing with continuous space. To incorporate the predictors for the continuous model, we would need a statistical sample from each potential location's geographic coordinates, which is clearly not possible over a continuous landscape where any location could serve as the ancestral node's origin. To address this problem, we propose a novel conceptual model for a quasi-continuous approach to phylogeography.

Proposed Method

We address the current gap between discrete and continuous phylogeography by introducing a quasi-continuous model in which additional discrete locations are added to the original observed locations of sequences. For the set K of n observed discrete locations where $n \in \mathbb{N}$, the user may specify an amount of new nodes, τ, where $\tau \in \mathbb{N}$ including 0, to be added for each n_k where $n_k \in K$ and $k \in [1, 2, \ldots, n]$. Let us define $\sigma_\tau = \tau$, $\sigma_{\tau\text{-}1} = \tau - 1, \ldots, \sigma_{\tau\text{-}(\tau\text{-}1)} = 1$ to represent the current value of τ as the algorithm proceeds and $i = 0$ to represent the iteration of the algorithm. That is, i increases in increments of 1 from 0 to $\tau - 1$ while σ decreases in increments of 1 from τ to 1. From each of the first new nodes, $\tau_{k0\tau}$, there will be $\sigma_{\tau\text{-}1} = \tau - 1$ new nodes, $\tau_{k1(\tau\text{-}1)}$. Each $\tau_{k1(\tau\text{-}1)}$ has $\sigma_{\tau\text{-}2} = \tau - 2$ new nodes, and this process continues until $\tau = 1$ when there will be no more new nodes to add. That is, each new node $\tau_{ki\sigma}$ will have $\sigma_\tau = \tau - (i + 1)$ new nodes. The $\tau_{k0\tau}$ nodes will be dispersed equally about n_k at an angle $\theta_{k0\tau}$ where $\theta_{k0\tau} = [2\pi / (\tau + 1)]$ radians relative to the vector $v_{n_k n_j}$ connecting n_k and its nearest neighbor n_j where $n_j \in K$. The distance of each $\tau_{k0\tau}$ node from n_k will be a length $\alpha_{k0\tau}$ where $\alpha_{k0\tau}$ is half the distance between observed locations n_k and n_j, that is $\alpha_{k0\tau} = \| v_{n_k n_j} \| / 2$. The $\tau_{k1(\tau\text{-}1)}$ nodes will be dispersed about $\tau_{k0\tau}$ in a similar manner such that the distance $\alpha_{k1(\tau\text{-}1)}$ is half the distance between $\tau_{k0\tau}$ and n_k and at an angle $\theta_{k1(\tau\text{-}1)}$ where $\theta_{k1(\tau\text{-}1)} = [2\pi / ((\tau - 1) + 1)]$ radians relative to the vector $v_{\tau_{k0\tau} n_k}$ connecting n_k and $\tau_{k0\tau}$.

This dispersal pattern will continue for all i, σ, and n_k such that the angle and distance of each node to be added distributed by (3), (4), and (5). Box 1 shows the pseudocode of this algorithm and one example is shown in Figure 1.

$$\theta_{ki\sigma_\tau} = 2\pi \,/\, (\sigma_\tau + 1) \tag{3}$$

$$\alpha_{ki\sigma_\tau} = ||v_{n_k n_j}|| \,/\, 2 \qquad \text{when i = 0, } \sigma = \tau \tag{4}$$

$$\alpha_{ki\sigma_\tau} = \alpha_{ki\sigma_\tau} \,/\, 2 \qquad \text{when i > 1, } \sigma < \tau \tag{5}$$

Prompt user to enter τ, the desired number of new locations to be added per observed discrete location n_k in K
For each observed discrete state n_k in K
  Determine the nearest neighbor n_j
  For $\sigma = \tau$ to $\sigma = 1$
    Draw σ new nodes, p, from n_k at a distance α_p from n_k where $\alpha_p = \| n_k - n_j \| / 2$
    Space each node p at $\theta_p = [2\pi / (\sigma+1)]$ radians from each other p relative to a vector from n_k to n_j
      For m = σ – 1 to m = 1
        Draw m new nodes, q, from each node p at a distance α_m where $\alpha_m = \alpha_p / 2$
        Space each node q at $\theta_q = (2\pi / m+1)$ radians from each other q relative to a vector from p to n_k
        m = m– 1
  $\sigma = \sigma - 1$
  Move to the next observed discrete state

Box 1. Pseudocode for algorithm to create new locations for each node n_k in the set of original locations K.

In this algorithm, the distance between nodes will quickly decrease at the nodes added during the last several iterations. This will result in a high density of nodes near the outermost locations (high values of i) but the continuous revolution of angles from outer node to outer node and halving of the distance as ensures that the additional nodes will not occur at the same geographic location. It is also important to note that in this method, the total number of locations, ϕ, rapidly increases as τ and K increase. Table 1 demonstrates this trend, which is summarized by (6) and (7).

$$\phi\,(\tau, \mathrm{K}) = \mathrm{K}\,(\tau * \phi\,(\tau - 1, 1) + 1) \tag{6}$$

$$\phi\,(0, 1) = 1 \tag{7}$$

Here, (7) represents the base case for any (τ, K). This is intuitive, as with just one discrete state and zero additional locations to disperse about n_k there remains exactly one node. This base case, with just one discretized location, cannot provide any phylogeographic insight but each observed location n_k in K will scale on that base case and τ.

Results

Since this model is conceptual, it has yet to be implemented with phylogeographic software packages or tested for accuracy in Bayesian inference with a real set of discrete sites. Instead we provide a visualization of a specific application of this algorithm (Figure 1) and demonstrate the expansion of the observed sampling locations (Table 1) in our proposed novel quasi-continuous model.

Case	τ	$\phi \mid \tau$, K = 1	$\phi \mid \tau$, K = 2	$\phi \mid \tau$, K = 3	$\phi \mid \tau$, K = 4	$\phi \mid \tau$, K = 5	$\phi \mid \tau$, K = 6
1	0	1	2	3	4	5	6
2	1	2	4	6	8	10	12
3	2	5	10	15	20	25	30
4	3	16	32	48	64	80	96
5	4	65	130	195	260	325	390
6	5	326	652	978	1,304	1,630	1,956
7	6	1,957	3,914	5,871	7,828	9,785	11,742
8	7	13,700	27,400	41,100	54,800	68,500	82,200
9	8	109,601	219,202	328,803	438,404	548,005	657,606
10	9	986,410	1,972,820	2,959,230	3,945,640	4,932,050	5,918,460

Table 1. Numerical summary of the first 10 cases of additional locations, τ, to add to each state k in the original set of discrete locations K. Here ϕ represents the total number of locations given τ and set K.

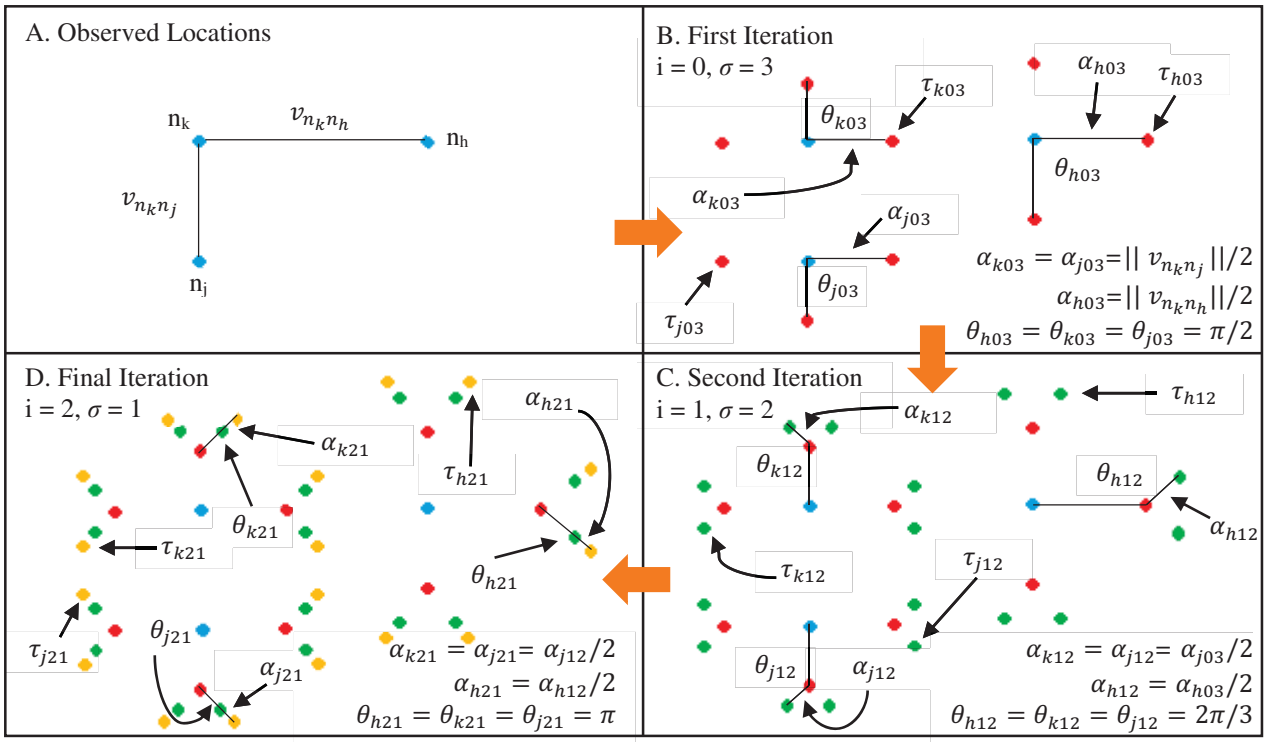


Figure 1. A step-by-step visual representation of the algorithm in Box 1 on a network of K = 3 observed discrete sampling locations with $\tau = 3$. A) The three observed sampling locations (n_h, n_j, n_k) are shown as blue circles and the corresponding shortest vectors are $v_{n_k n_j}$ and $v_{n_k n_h}$. B) Each observed location is given $\sigma = 3$ nodes (τ_{h03}, τ_{j03}, τ_{k03}) shown as red circles. These nodes are distributed by (3) and (4). C) Each red node is given $\sigma = 2$ nodes (τ_{h12}, τ_{j12}, τ_{k12}) shown as green circles, distributed by (3) and (5). D) Each green circle is given $\sigma = 1$ node, (τ_{h21}, τ_{j21}, τ_{k21}) shown as yellow circles, distributed by (3) and (5). At this point there are no new nodes to add and the algorithm exits. There are $\phi(3, 3) = 48$ total nodes in the new set by (6) and (7). The distances and angles between nodes are shown by α_{xyz} and θ_{xyz}, respectively, where x is the node (h, j, k), y is the i^{th} step in the iteration, and z is the count of σ nodes added during the step. Note that all α and θ values are equal for each node for each step in the algorithm.

Discussion

Our novel quasi-continuous model allows us to utilize the GLM for spatiotemporal hypothesis testing outside of a traditional Bayesian discrete setting. For each new location, the corresponding coordinate pair can mapped to see which discretized location it would fall in under the GLM. The predictor data for that location can then be that of the defined initial observed locations as seen in Lemey et al.[12] and Magee et al.[16] and be tested for Bayes factor support via (2). As previously mentioned, it is likely that as we reach the nodes for the smaller values of σ they will all fall in the same discretized location defined by the model because of their decreasing separation; however this will not be constraining for studies across a wider area where discretized locations are whole countries or global regions. With the increased locations that we have introduced in this quasi-continuous model, we will also be able to eliminate poorly reconstructed trees from consideration if an ancestral node lies in an unlikely geographic location such as an ocean, mountain range, or uninhabited desert. This increases the reliability of inferred ancestral states produced in phylogeographic software packages such as BEAST.

Incorporating landscape heterogeneity into a phylogeographic framework would undoubtedly yield dividends in accuracy, confidence, and reliability among inferred results. This challenging task is aided by publically available software and services such as Google Earth that can provide detailed information on global terrain. Although this integration has yet to be achieved, the impact on public health would be immediate as it would provide insight on the true origins of viral diffusion. It will also allow a more focused analysis on the local predictors associated with the dispersal of these viruses which could help identify the most plausible disease drivers via the GLM. Furthermore, data on climate, agriculture, livestock, and population demographics are becoming increasingly available via sources such as the National Oceanic and Atmospheric Administration and Food and Agricultural Organization of the United Nations. Reliable data sources such as these can provide the necessary inputs for the GLMs and the ability of the GLM

to identify drivers of viral diffusion increases as the number of predictor variables increases.

There are limitations with this model, including the lack of hypothesis testing on actual data that would demonstrate how our model can be visualized, analyzed, and interpreted. In addition, the model will be computationally intensive for larger sets of observed locations and larger values of τ. The eigen decomposition of the rate matrix between the increased number of locations may cause problems with these software packages. For a large number of discrete, sparse locations a continuous model is generally the better option, but due diligence should be performed to analyze the performance of this conceptual model in our quest for the incorporation of the GLM into continuous space.

Although this quasi-continuous model does not quite complete the desired task of integrating a GLM within a continuous Bayesian phylogeographic model but does improve upon the established discrete GLM by including more nodes at the request of the end user. Future work will include incorporating these concepts into the BEAST framework such that they become accessible to users, allowing a specific value of τ for each observed sampling location, and eliminating added nodes from consideration prior to hypothesis testing if they fall in an unlikely location. In addition, we will be able to statistically analyze the effect of τ on the model and identify an optimal value for computer performance and Bayesian inference. Once completed, we will have the capability to test this model, identify flaws, and strengthen the algorithm to expand and improve the field of phylogeography.

Acknowledgements
The project described was supported by award number R00LM009825 from the National Library of Medicine to Matthew Scotch. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine, the National Institutes of Health. The authors would like to thank Philippe Lemey, Ph.D., for his suggestions regarding a GLM approach for a large number of observed sampling locations.

References
1. Scotch M, Mei C, Makonnen Y, et al. Phylogeography of influenza a H5N1 clade 2.2.1.1 in Egypt. BMC Genomics. 2013;14(1):871.
2. May FJ, Davis CT, Tesh RB, Barrett ADT. Phylogeography of West Nile Virus: from the Cradle of Evolution in Africa to Eurasia, Australia, and the Americas. Journal of Virology. 2011 March 15, 2011;85(6):2964-74.
3. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography Takes a Relaxed Random Walk in Continuous Space and Time. Molecular Biology and Evolution. 2010 August 1, 2010;27(8):1877-85.
4. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. PLoS Comput Biol. 2009;5(9):e1000520.
5. Lemmon AR, Lemmon EM. A Likelihood Framework for Estimating Phylogeographic History on a Continuous Landscape. Systematic Biology. 2008 August 1, 2008;57(4):544-61.
6. Krauss H. Zoonoses: Infectious Diseases Transmissible from Animals to Humans: ASM Press; 2003.
7. Chen Y, Liu T, Cai L, Du H, Li M. A One-Step RT-PCR Array for Detection and Differentiation of Zoonotic Influenza Viruses H5N1, H9N2, and H1N1. Journal of Clinical Laboratory Analysis. 2013;27(6):450-60.
8. He D, Dushoff J, Eftimie R, Earn DJD. Patterns of spread of influenza A in Canada. Proceedings of the Royal Society B: Biological Sciences. 2013 November 7, 2013;280(1770).
9. Loth L, Gilbert M, Wu J, Czarnecki C, Hidayat M, Xiao X. Identifying risk factors of highly pathogenic avian influenza (H5N1 subtype) in Indonesia. Preventive Veterinary Medicine. 2011 10/1/;102(1):50-8.
10. Pfeiffer DU, Minh PQ, Martin V, Epprecht M, Otte MJ. An analysis of the spatial and temporal patterns of highly pathogenic avian influenza occurrence in Vietnam using national surveillance data. The Veterinary Journal. 2007 9//;174(2):302-9.
11. Gilbert M, Xiao X, Pfeiffer DU, et al. Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. Proceedings of the National Academy of Sciences. 2008 March 25, 2008;105(12):4769-74.
12. Lemey P, Rambaut A, Bedford T, et al. The seasonal flight of influenza: a unified framework for spatiotemporal hypothesis testing. arXiv:12105877v1. 2012.
13. McCullagh P. Generalized linear models. European Journal of Operational Research. 1984 6//;16(3):285-92.
14. Suchard MA, Weiss RE, Sinsheimer JS. Models for Estimating Bayes Factors with Applications to Phylogeny and Tests of Monophyly. Biometrics. 2005;61(3):665-73.
15. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug;29(8):1969-73.
16. Magee D, Beard R, Suchard MA, Lemey P, Scotch M. Combining phylogeography and spatial epidemiology to uncover predictors of H5N1 influenza A virus diffusion. Archives of Virology. 2014 Oct 30.

Mining Electronic Health Records using Linked Data

David J. Odgers, MS, Michel Dumontier, PhD
Stanford Center for Biomedical Informatics Research, Stanford University, Stanford, CA

Abstract

Meaningful Use guidelines have pushed the United States Healthcare System to adopt electronic health record systems (EHRs) at an unprecedented rate. Hospitals and medical centers are providing access to clinical data via clinical *data warehouses such as i2b2, or Stanford's STRIDE database. In order to realize the potential of using these data* for translational research, clinical data warehouses must be interoperable with standardized health terminologies, biomedical ontologies, and growing networks of Linked Open Data such as Bio2RDF. Applying the principles of Linked Data, we transformed a de-identified version of the STRIDE into a semantic clinical data warehouse containing visits, labs, diagnoses, prescriptions, and annotated clinical notes. We demonstrate the utility of this system though basic cohort selection, phenotypic profiling, and identification of disease genes. This work is significant in that it demonstrates the feasibility of using semantic web technologies to directly exploit existing biomedical ontologies and Linked Open Data.

Introduction

Driven by Meaningful Use [1,2] guidelines, the United States Healthcare System has experienced a widespread adoption of Electronic Health Records (EHR). This has provided biomedical researchers with data warehouses which promote breakthrough research and lead to enhanced patient care. Data within EHR systems, typically described with standard health terminologies (SNOMED-CT, ICD9, RxNORM, LOINC), can be used to identify and profile patient cohorts cross-sectionally and longitudinally to investigate disease progression, drug safety and efficacy, genotype and phenotype variation and laboratory measurement anomalies down to the biochemical level[3,4]. However, with nearly 400 open biomedical ontologies available in BioPortal [5], and dozens of biomedical datasets being made available as part of the Bio2RDF network of Linked Open Data (LOD), there is a salient opportunity to integrate clinical and biomedical data together to better understand patient populations and to uncover associations of biomedical interest. Towards these long term goals, we transformed Stanford's STRIDE clinical data warehouse into an integrated, semantic knowledge base that uses ontologies to bridge the gap between clinical and biomedical data. We demonstrate the utility of our system through patient cohort selection, phenotypic profiling, with an eye to future studies focused on drug repositioning, combination therapies, and exploring the complex interplay of genetics, biochemistry, and lifestyle. Our system has the potential to accelerate translational research from the bedside to the bench via efficient approaches for knowledge discovery that are required in the complex interrogation of biomedical data.

Background

The Semantic Web is an effort to provide machine understandable data that builds on web technology and provides standardized languages and interfaces. Groundbreaking work by Prathak [6,7,8,9,10], et al demonstrated how semantic web technologies could be used to integrate the Mayo Clinic EHR with existing biomedical data to investigate genetic factors and comorbidities associated with Type 2 Diabetes Mellitus (T2DM), explore genotype/phenotype associations from BioBank tissue samples for T2DM, identify drug-drug-interactions using clinical diagnosis and prescription, and execute "on-the-fly" cohort selection that links with up to 12 different prescription drug datasets. With the growing availability of public biomedical ontologies and datasets such as disease, drug, and phenotype ontologies along with drug product labels, clinical trials, spontaneous adverse events reports, the value proposition continues to grow for access to clinical data warehouses that directly interoperate with these and future data resources.

Bio2RDF. Bio2RDF is the largest open-source, semantic web repository of life science data on the internet, containing ~11 billion triples across 35 datasets. Bio2RDF includes data of biomedical and clinical interest including chemicals, genes, drugs, drug targets, drug indications, diseases, bioassays, genotype-phenotype data, pharmacogenomic data, clinical trials, and drug product labels. The Bio2RDF network provides a number of ways to query from EHR data directly into high quality basic biology resources on the web [11]. A key part of the success of this project is based on the normalization of data identifiers as template-based Uniform Resource Identifiers (URIs). Each

RDF dataset is loaded into an RDF-specific data store (aka triple store) which enables query answering using the SPARQL query language over the web-friendly HTTP protocol [14,15,16].

STRIDE. STRIDE is a central repository for EHR data from the Lucile Packard Children's Hospital and Stanford Hospital and Clinics. The subset of EHR data that we have used for this system is generated from 18 years of data (1994-2011), 1.8 million patients, 19 million encounters, 35 million coded ICD9 diagnosis and more than 11 million unstructured clinical notes which are a combination of pathology, radiology and transcription reports. The dataset includes both inpatient and outpatient notes that include radiology, pathology, and transcription reports. Figure 1 demonstrates the workflow to annotate clinical notes in STRIDE [14,15,16,17]. Additionally, prescription and detailed visit information with structured International Classification of Disease (ICD9) codes and Current Procedural Terminology (CPT) codes are available.

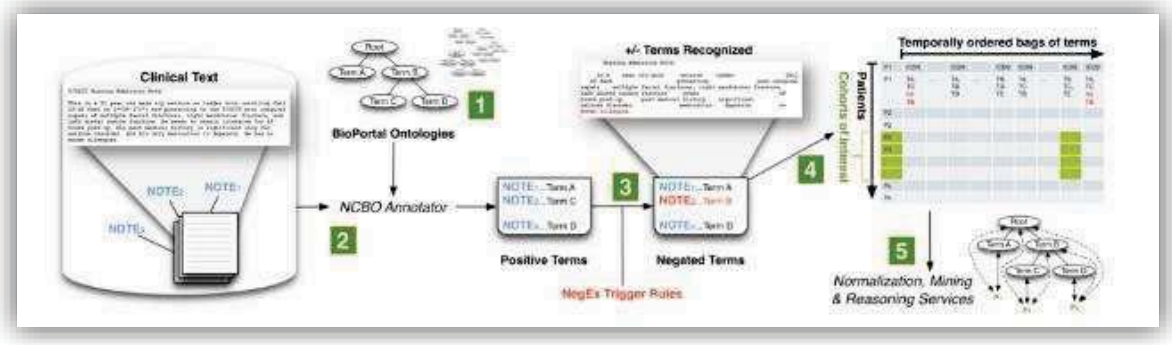


Figure 1. Annotator Workflow for Clinical Note annotation.

The NCBO Annotator Workflow extracts terms from the clinical notes of patients: (1) Create a lexicon of over 2.8-million terms from the NCBO BioPortal library. (2) Use the NCBO Annotator to rapidly find those terms in clinical notes—which are called annotations. (3) Apply NegEx trigger rules to separate negated terms. (4) Compile terms (both positive and negative) into a temporally ordered series of sets for each patient and combine them with coded and structured data when possible. (5) Reason over the structure of the ontologies to normalize and to aggregate terms for further analysis [17].

Methods

Extract, Transform, Load STRIDE2RDF Dataset

We developed PHP and Python scripts to transform a de-identified extract of the STRIDE CDW into RDF using Bio2RDF guidelines [14,15,16]. The RDF was then loaded into an instance of the Virtuoso 7.1.0 open source community edition database with an access-limited, federated query enabled, SPARQL endpoint. In this way, the STRIDE2RDF endpoint can execute queries on Bio2RDF datasets to complete queries of interest. Our choice to perform an ETL into a persistent storage sharply contrasts with that of Mayo Clinic's SHARPn architecture that uses an SQL-based RDF virtual views [7,13,21]. Performance of triple stores has vastly improved over the past few years, and persistent stores now offer significant speed, efficiency and flexibility.

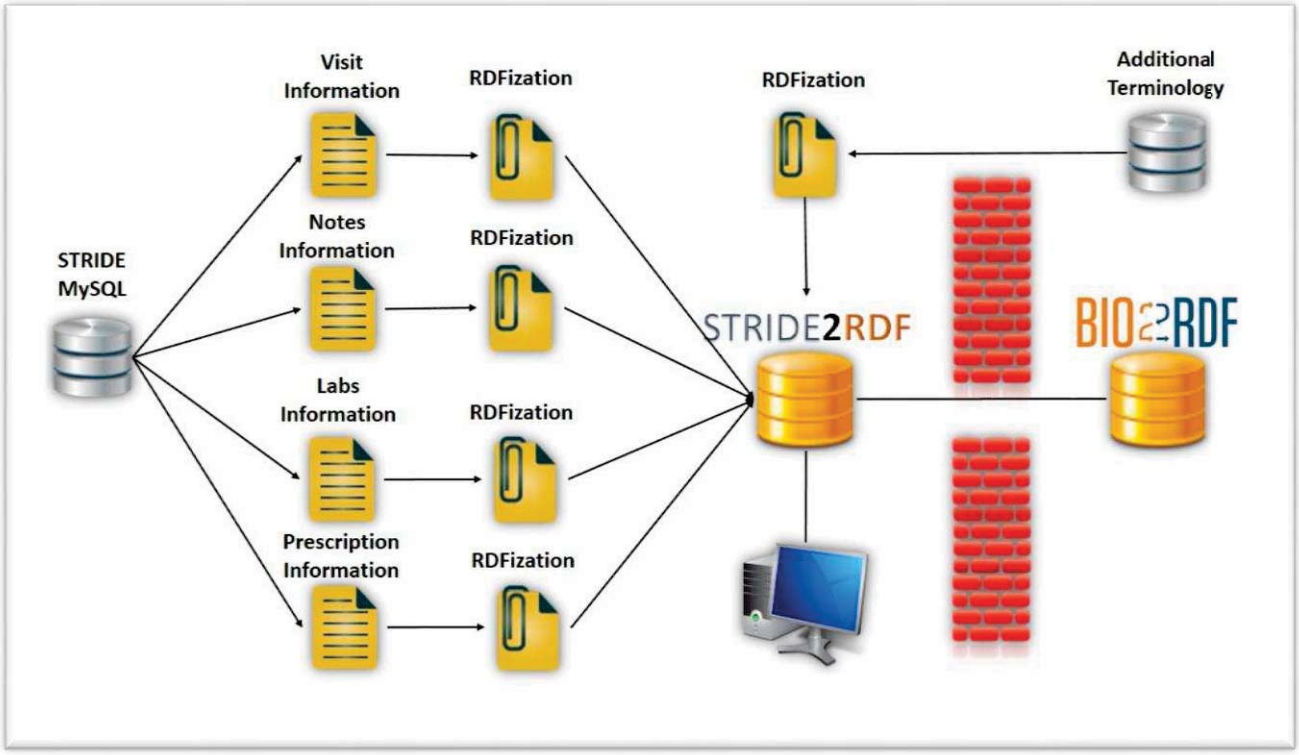


Figure 2. STRIDE2RDF Architecture.

STRIDE2RDF architecture comprises of a restricted access endpoint to limit access to only authorized users. Our work is considered non-human subjects by the Stanford IRB. Federated queries can be made by authorized users to select external content required to complete query. The federated query will be aggregated behind a firewall, to ensure that clinical data is never exposed to an open network. . The ETL process extracted prescription, lab, note and visit information from the STRIDE SQL database and terminology from publicly available ontologies, performed an RDFization transformation and loaded normalized RDF triples into a secure Virtuoso triple store.

Results

We evaluated our system with a set of 10 questions, of which 3 are presented below. Our queries were executed from an end user console, behind a firewall, as depicted figure 2. These exemplar queries demonstrate that we can build patient cohorts using selected attributes - coded diagnoses and clinical note annotations - and connect into selected biomedical terminologies and Linked Datasets - OMIM [22], SIDER [23]. The questions relate to Mucopolysaccharidosis, a group of rare metabolic disorders caused by dysfunction in lysosomal storage enzymes. The first question uses diagnoses associated with patient visits to identify other diseases that are experienced by the patient throughout their lifetime. The second question uses Bio2RDF's version of OMIM to identify disease genes that are associated with the co-morbid diseases. The third question uses ICD9, RxNorm, and SIDER to identify known drug side effects that are experienced by Mucopolysaccharidosis patients taking Trometamine for metabolic acidosis.

	Question	Dataset used
1	What co-morbidities are most often found in patients that suffer from Mucopolysaccharidosis?	STRIDE2RDF, ICD9
2	What disease genes are associated with Mucopolysaccharidosis co-morbidities?	STRIDE2RDF, ICD9, OMIM
3	Which adverse events experienced by Mucopolysaccharadosis patients taking Tromethamine are associated with this drug?	STRIDE2RDF, ICD9, RxNORM, SIDER

Figure 3. Exploratory queries for the STRIDE2RDF graph. Results and SPARQL queries can be found online at http://tinyurl.com/l7f8yjj.

Discussion

In this paper we describe STRIDE2RDF, a semantic clinical data warehouse for constructing patient cohorts and undertaking translational research by linking out to external biomedical datasets through standard health care terminologies. With increasing amounts of EHR data coupled with growing amounts of biomedical ontologies and biological datasets, our work pushes the boundaries of ubiquitous data access in support of translational research. It's worth noting that hundreds of additional datasets such as data.gov, DBpedia, World FactBook, Semantic Tweet, are available on the Semantic Web, which could be used to extend studies into altogether new areas with minimal effort.

STRIDE2RDF represents a machine and human interpretable, formal knowledge representation that is much more expressive than a standard SQL clinical database. Our representation is amenable to conjunctive query answering using federated SPARQL queries. This enables access to deductive reasoning using the expert knowledge contained in OWL ontologies (primarily transitive closure of any relation), and to simultaneously query outside resources through service calls within the query itself. Our work paves the way for more sophisticated analyses such as using OWL ontologies to check the consistency of the knowledge base [24] and finding new associations between linked entities [25].

Limitations

While promising, this proof of concept requires more development and optimization to realize its full potential. The primary limitations of this platform are i) the limited number of outwards links, ii) performance of federated queries, iii) URI mismatches, iv) scalability, v) dirty clinical data. Outwards links are currently restricted by the use of standard health care dictionaries (SNOMED-CT, ICD9, RxNORM, LOINC). Our goal is to use mappings in the UMLS as a way to traverse from these terminologies into other public biomedical ontologies and resources. For instance, Orphanet provides additional characterization of rare diseases, such as phenotypes and their frequency which may be useful in clinical decision support. Our use of a federated query was found to be sub-optimal for queries involving thousands of concept joins with external data. However, since these external data are made available in RDF, we anticipate substantial performance gains when loaded into a local triple store. While the data in the Bio2RDF network are intrinsically connected by virtue of steadfast adherence to a common URI pattern (e.g. http://bio2rdf.org/prefix:identifier, where prefix is a globally unique shortname for the dataset), these do not always align with external resources. With the advent of the identifiers.org SPARQL service for automatically resolving Bio2RDF identifiers with other identifier systems, URI mismatches will now be less of an issue [26].

Scalability and accessibility have been ongoing concerns for semantic web technologies. In the absence of efficient query planners and highly optimized implementations, naive queries may result in poor performance. However, the needs for more expressive queries capable of incorporating the semantics of hierarchies and terminology mappings necessitates more sophisticated solutions. As the field matures and simpler and more effective solutions such as Linked Data Fragments [27] become more widely used, the burden of using these technologies diminishes. It's worth noting that RDF can be serialized in a variety of formats including XML and JSON, thereby providing a concrete mechanism to make it available to a wider community of application developers.

Future Work

We plan to address the limitations described above and begin to explore the EHR as a starting point for a number of studies including, but not be limited to, retrospective clinical studies, biomarker discovery, patient stratification, drug repurposing, and pharmacogenomics.

Contributions

David Odgers (Student) contributed to the ETL of the clinical data, query generation, background research, and the bulk of the paper authorship. Michel Dumontier (Primary Adviser) contributed to the RDFization of the clinical data, data integration, SPARQL endpoint integration and HIPAA compliance concerns.

References

1. Blumenthal, David, and Marilyn Tavenner. The "meaningful use" regulation for electronic health records. New England Journal of Medicine. 2010:363.6:501-504.

2. Jha, Ashish K. Meaningful use of electronic health records: the road ahead. JAMA. 2010:304.15:1709-1710.
3. Jensen, Peter B., Lars J. Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012:13.6:395-405.
4. Luciano, Joanne S., et al. The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. J. Biomedical Semantics. 2011:2.S-2: S1.
5. Bioportal Homepage [homepage on the Internet]. Stanford University. [cited 13 Aug 2014]. Available from: http://bioportal.bioontology.org/.
6. Pathak, Jyotishman, et al. Applying semantic web technologies for phenome-wide scan using an electronic health record linked biobank. J. Biomedical Semantics. 2012: 3:10.
7. Pathak, Jyotishman, et al. Mining the human phenome using semantic web technologies: a case study for type 2 diabetes. AMIA Annual Symposium Proceedings 2012. American Medical Informatics Association. 2012.
8. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Using Linked Data for Mining Drug-Drug Interactions in Electronic Health Records. Studies in health technology and informatics. 2013:192: 682.
9. Pathak, Jyotishman, et al. Validation and discovery of genotype-phenotype associations in chronic diseases using linked data. Studies in health technology and informatics. 2011:180:549-553.
10. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Using semantic web technologies for cohort identification from electronic health records for clinical research. AMIA Summits on Translational Science Proceedings 2012. 2012:10.
11. Bio2rdf.org [homepage on the Internet]. Carleton University. [cited 13 Aug 2014]. Available from: https://Bio2RDF.org.
12. W3C Website [Internet]. Semantic Web Standards. [cited 14 Aug 2014]. Available from: http://www.w3.org/standards/semanticweb/.
13. Pathak, Jyotishman, Richard C. Kiefer, and Christopher G. Chute. Mining anti-coagulant drug-drug interactions from electronic health records Using Linked Data. Data Integration in the Life Sciences. 2013:128-140.
14. Nolin, Marc-Alexandre, et al. Bio2RDF network of linked data. Semantic Web Challenge. International Semantic Web Conference (ISWC 2008). 2008.
15. Cruz-Toledo, José, Alison Callahan, and Michel Dumontier. Bio2RDF: linked data for the life sciences. 2013.
16. Callahan, Alison, José Cruz-Toledo, and Michel Dumontier. Ontology-based querying with Bio2RDF's linked open data. J. Biomedical Semantics. 2013:4.S-1: S1.
17. LePendu, Paea, et al. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Summits on Translational Science Proceedings 2012. 2012: 63.
18. LePendu, Paea, et al. Pharmacovigilance using clinical notes. Clinical pharmacology & therapeutics. 2013:93.6: 547-555.
19. LePendu, Paea, et al. Case studies in making sense of clinical text. Proceedings of the BioLINK SIG 2013. 2013.
20. Iyer, Srinivasan V., et al. "Mining clinical text for signals of adverse drug-drug interactions. Journal of the American Medical Informatics Association. 2014:21.2:353-362.
21. Rea, Susan, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project. Journal of biomedical informatics. 2012:45.4:763-771.
22. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 14 Dec 2014. World Wide Web URL: http://omim.org/.
23. Kuhn, M., et al. A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol. 2010;6:343. Epub 2010 Jan 19.
24. Hoehndorf, Robert., et al. Integrating systems biology models and biomedical ontologies. BMC systems biology. 2011:5.1:124.S.
25. Hoehndorf, Robert, Michel Dumontier, and Georgios V. Gkoutos. Identifying aberrant pathways through integrated analysis of knowledge in pharmacogenomics. Bioinformatics. 2012:28.16 :2169-2175.
26. Identifiers.org Homepage [homepage on the Internet]. EMBL-EBI. [cited 25 Sept 2014]. Available from: http://identifiers.org/services/sparql.
27. Verborgh, R., et al. Querying datasets on the Web with high availability. The Semantic Web–ISWC 2014. 2014. 180-196.

Concept Modeling-based Drug Repositioning

Jagadeesh Patchala[1] and Anil G Jegga[1,2,3]
[1]Department of Computer Science, [2]Division of Biomedical Informatics, [3]Department of Pediatrics, Cincinnati Children's Hospital and Medical Center, University of Cincinnati, Cincinnati, Ohio, USA

Abstract
Our hypothesis is that drugs and diseases sharing similar biomedical and genomic concepts are likely to be related, and thus repositioning opportunities can be identified by ranking drugs based on the incidence of shared similar concepts with diseases and vice versa. To test this, we constructed a probabilistic topic model based on the Unified Medical Language System (UMLS) concepts that appear in the disease and drug related abstracts in MEDLINE. The resulting probabilistic topic associations were used to measure the similarity between disease and drugs. The success of the proposed model is evaluated using a set of repositioned drugs, and comparing a drug's ranking based on its similarity to the original and new indication. We then applied the model to rare disorders and compared them to all approved drugs to facilitate "systematically serendipitous" discovery of relationships between rare diseases and existing drugs, some of which could be potential repositioning candidates.

Introduction
Drug repositioning is the process of developing new indications for existing drugs or biologics. Maximizing the indications potential and revenue from drugs that are already marketed offers a new take on the famous mantra of the Nobel Prize-winning pharmacologist, Sir James Black, "*The most fruitful basis for the discovery of a new drug is to start with an old drug*". Rational design of drug mixtures however poses formidable challenges because often the details of *in vivo* cell regulation and pathway interactions and mechanisms underlying genetic pathway regulation are obscure. Thus, several of the repositioned drugs are discovered serendipitously in the form of unexpected findings during late phase clinical studies. One of the reasons that the connection between drug candidates and their potential new indications could not be identified earlier is that the underlying mechanism associating them is either very intricate and unknown or dispersed and buried in a sea of information. Drug repositioning is predominantly dependent on two principles: i) the "promiscuous" nature of the drug and ii) targets relevant to a specific disease or pathway may also be critical for other diseases or pathways[1,2]. The latter may be represented as a shared gene or biomedical concept between a disease-disease, drug-drug, or a disease-drug. Based on this principle, some computational approaches have been developed and applied to identify drug repositioning candidates ranging from mapping gene expression profiles with drug response profiles to side-effect based similarities[3-8].

The topic model is a state-of-the-art Bayesian model for extracting semantic structure from document collections[9]. It automatically learns a set of thematic topics (lists of words or "bag of words") that describe a document collection, and assigns the topics to each of the documents in the collection with a probability value. Topic models have recently retained a lot of attention and have been used to address various issues (e.g., drug repositioning[10], word sense disambiguation in the clinical domain[11], gene-drug relationship extraction from literature[12], etc.). As a variation of classic "bag-of-words" approach, we use a "bag of concepts" approach. We first employ the UMLS Metathesaurus to identify biomedical concepts and construct a probabilistic topic model based on the concepts that appear in the disease and drug related abstracts. The resulting probabilistic topic associations are used to measure the similarity between disease and drugs and identify drug repositioning candidates (Fig. 1).

Methods
MEDLINE Abstract collection
Disease and drug-related abstracts were extracted from MEDLINE using NCBI's E-Utilities feature[13]. We created PubMed queries (using disease or drug names along with the MeSH field tag, if available) that returned respective list of articles (ranging from 100 to 10000). For topic modeling purposes, we only used PubMed search results that contained abstracts. From the collected sets of abstracts, we randomly selected 500 abstracts with mapped concepts (see section Concept Mapping) for topic modeling (Fig. 1). For validation purposes, we selected 11 disease-drug pairs representing known and candidate repositioned drugs (e.g., ropinirole-Parkinson's disease and ropinirole-Restless legs syndrome) and downloaded all the abstracts related to the disease and drug. Abstracts that cited both disease and drug are excluded from topic modeling input to avoid the over-fitting of our model to any particular drug or disease. In other words, if an abstract cites both the disease and drug from select disease-drug pairs (e.g. abstracts citing both ropinirole and Parkinson's disease), it was not used to generate the topics. As our test set, we collected the list of 1704 approved drugs from the DrugBank[14] and six rare diseases. For each of these diseases and

drugs we compiled the list of published articles and randomly selected 500 abstracts for each, at a time, for the analysis. We removed 10 drugs (from total 1704) from our drug data set because at the time of this analysis, each of these drugs had fewer than 50 publications. Our final dataset thus comprised 1694 drugs and 6 rare diseases resulting in about 850K (1700 drugs/diseases * 500 abstracts) abstracts. Our goal is to rank the 1694 drugs based on their likelihood as drug repositioning candidates for each of the selected six rare diseases as measured by their similarity to the six rare diseases. In each of the runs (total 10 runs for each disease), we changed the 500 abstracts for the rare disease and drugs and recorded the top ranked drugs.

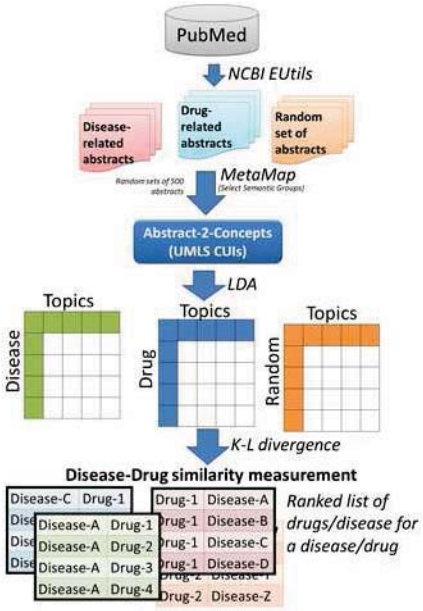


Fig. 1: Schematic representation of overall workflow. Drug and disease-related abstracts are Metamapped to generate a list of biomedical and genomic CUIs from UMLS for each drug and disease. Topic modeling is then applied followed by statistical analysis to assess the similarity between disease and drug.

Concept Mapping

To map the downloaded disease and drug related abstracts to concepts from the UMLS Metathesaurus, we used MetaMap[15] with semantic types restricted to five semantic groups, namely, Anatomy, Chemicals and Drugs, Disorders, Genes and Molecular Sequences, and Physiology. MetaMap returns the list of candidate mappings (along with their score) and all of the MetaMap identified concepts from the five select semantic groups with a score of >350 were used for topic modeling. We used the Concept Unique Identifiers (CUIs) as input instead of concept terms to avoid redundancy and increase the specificity of the model (Fig. 1).

Topic Modeling

For a document d, $\theta(d)= P(t)$ stands for the multinomial distribution over topics. Let $P(w|t)$ be the probability distribution over words w given topic t. Then, following process generates the words in the document d, $P(w_i) = \sum_{j=1}^{t} P(\frac{Wi}{Ti} = j)P(Ti = j)$,where T is the number of topics. We used MALLET (MAchine Learning for LanguagE Toolkit)[16] a JAVA-based package to build our topic model. To determine the number of inherent topics in our data set, we first started with a topic size of 25 and calculated the log likelihood value for the trained model. The log likelihood value indicates how well the topic model fits with the data. By keeping the other parameters constant, we increased the topic size in multiples of 25 till 500 topics and calculate the log likelihood values for each scenario. The best number of topics is the one with the highest log likelihood value which was between 175-200 topics in this case. We therefore set the number of topics at 200.

Disease-drug distance assessment

We use Kullback-Leibler (KL) divergence[17] to compute the differences between the topic distributions in the selected disease and drug profiles. Given two uncertain objects P and Q and their corresponding probability distributions, KL divergence measures the similarity between two probability distributions and represents the information lost when Q is used to represent P. It is calculated as: $D_{KL}(P \parallel Q) = \sum_i P_i \log_2 (P_i/Q_i)$. Even though KL divergence is predominantly used to calculate the distance between two probability distributions, it is not a true metric as it is not symmetric. The KL divergence of P and Q is not equal to KL divergence of Q and P, unless P and Q are equal. In the current study, we therefore calculate the symmetric form of KL divergence, which is given by $D(P, Q)=D_{KL}(P \parallel Q)+D_{KL}(Q \parallel P)$. We use the intuitive idea that the drugs that are likely to be repurposed for a disease will have significantly small KL values when compared to the average of the drug distances to that disease. We capture this notion by imposing the condition that for a drug to be considered as repositioning candidate it should have a Z-score of -1.5 compared to the average. In other words, if a drug's divergence value is significantly smaller than the average divergence of all the drugs we consider it as a potential candidate for repositioning. We thus rank the drugs that have Z-score of -1.5 or lower according to their KL distance values and display them as the probable drugs that can be repositioned for that disease

Table 1: Selected examples of drugs with multiple indications

Drug	Indication-1	Indication-2
Formoterol	Asthma	Stuttering
Mitoxantrone	Multiple sclerosis	Prostate cancer
Modafinil	Narcolepsy	Bipolar disorder
Ropinirole	Parkinson's disease	Restless legs syndrome
SSRIs	Depression	Dysmorphic disorders
Terbutaline	Asthma	Preterm labor

Results

Validation

We select 6 examples of repositioned drugs (11 disease-drug pairs; 5 drugs with two indications each and one class of depression-related drugs as repositioning candidates for dysmorphology) to validate our approach (Table 1). The goal was to see how topic model-based approach will rank the drug against its multiple indications (i.e., drug-A vs. disease-1 and drug-A vs. disease-2). As described in Methods, we downloaded the drug and disease related abstracts, excluding abstracts, which cite both drug and disease. We mixed 9 random drug profiles with each disease-drug pair and calculated the rank of the original drug for the disease. We repeated this process 10 times for each disease-drug pair and calculated the accuracy, balanced accuracy, and precision as follows:

$$\text{Accuracy} = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives}$$

$$\text{Precision} = \frac{true\ positives}{true\ positives + false\ positives}$$

$$\text{Balanced accuracy} = \frac{sensitivity + specificity}{2}$$

For the validation sets, accuracy, balanced accuracy, and precision were 0.83, 0.75, and 0.32 respectively with an AUC of 0.74. The target drug was ranked at top 61% of the time.

Table 2: Top 15 ranked drugs for six rare diseases

Rank	Polycythaemia vera	Primary myelofibrosis	Dravet syndrome	Meningioma	Narcolepsy	Netherton syndrome
1	Pipobroman	Pipobroman	Riluzole	Lomustine	Zolpidem	Clocortolone
2	Ruxolitinib	Anagrelide	Alglucosidase alfa	Temozolomide	Cloxazolam	Amcinonide
3	Tofacitinib	Ruxolitinib	Ethosuximide	Dacarbazine	Ketazolam	Prednicarbate
4	Anagrelide	Oprelvekin	Creatine	Procarbazine	Zaleplon	Alclometasone
5	Eltrombopag	Hydroxyurea	Choline	Ifosfamide	Estazolam	Flurandrenolide
6	Uracil mustard	Pomalidomide	Valproic Acid	Altretamine	Camazepam	Diflorasone
7	Oprelvekin	Tofacitinib	Lamotrigine	Carmustine	Zopiclone	Fluocinonide
8	L-Tyrosine	Pegademase bovine	Clobazam	Mechlorethamine	Halazepam	Clobetasol propionate
9	Ginseng	Bortezomib	Zonisamide	Topotecan	Quazepam	Halobetasol Propionate
10	Hydroxyurea	Busulfan	Perampanel	Dexrazoxane	Chlordiazepoxide	Flumethasone Pivalate
11	Pegademase bovine	Thalidomide	Phenacemide	Etoposide	Bromazepam	Desoximetasone
12	Pomalidomide	Becaplermin	Topiramate	Daunorubicin	Triazolam	Monobenzone
13	Acetylsalicylic acid	Lenalidomide	Pilocarpine	Vincristine	Tofisopam	Desonide
14	Bortezomib	Fludarabine	Paramethadione	Vindesine	Delorazepam	Acitretin
15	Acenocoumarol	Eltrombopag	Trimethadione	Ethiodized oil	Clotiazepam	Betamethasone

New indication search – Drug Repositioning candidates for rare diseases

For each of the six rare disorders, we identified the nearest drug neighbors by calculating the KL distance between the rare disease and all of the 1694 drugs. We repeated this 10 times by changing the profile set of the rare diseases and recorded the number of times a specific drug was ranked among top 15 out of a total ten iterations. Table 2 enlists the top 15 ranked drugs for each of the 6 rare diseases. Literature search showed that most of the top ranked drugs could be related to their mapped respective rare diseases suggesting the utility of our approach in discovering drug repositioning candidates. In the following sections we discuss a few of our findings.

In case of polycythemia vera (PV), a rare bone marrow disease that leads to an abnormal increase in the number of blood cells, the top ranked drug in our analysis is pipobroman. There are several studies reporting the efficacy of pipobroman in PV[18,19]. Ruxolitinib ranked second for PV and third for primary myelofibrosis in our analysis. Ruxolitinib has been recently reported to provide clinical benefits in patients with advanced PV[20] and in primary myelofibrosis[21].

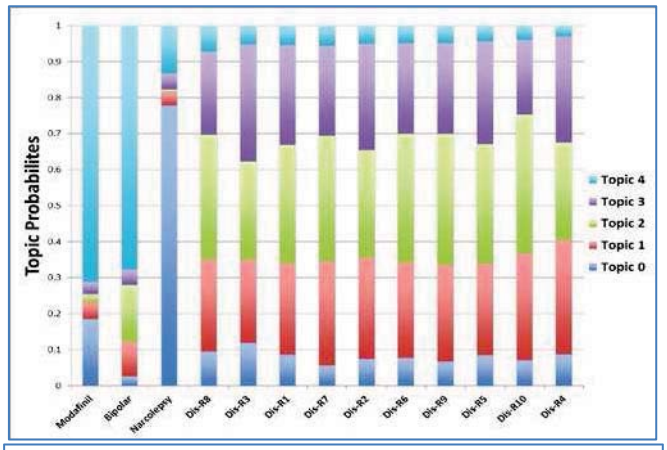


Fig. 2: Stacked bar chart showing the top five topic proportions found in modafinil (drug) and its two indications (bipolar disorder and narcolepsy) and ten random disease sets.

Dravet syndrome, a rare genetic epileptic encephalopathy, is primarily caused by mutations in the voltage-gated sodium channel *SCN1A* gene. The top ranked drug in our analysis for Dravet syndrome is riluzole, a sodium channel inhibitor. Although loss-of-function mutations are common in Dravet syndrome, a gain-of-function mutation[22] and duplications[23] in *SCN1A* have also been reported suggesting that sodium channel inhibitors like riluzole may be useful in such cases. Interestingly, riluzole was first developed as an anti-epileptic drug but is now used for treatment of amyotrophic lateral sclerosis[24]. The other top

ranked drugs for Dravet syndrome were various antiepileptic drugs. Likewise, for meningiomas, a diverse set of tumors arising from the meninges, among the top ranked drugs were several candidates that are currently investigated for various forms of brain tumors.

We also note that a high conceptual similarity between a disease and drug may not always suggest alternate indication but semantic relatedness or potential contraindication or even drug related side-effects. For example, in narcolepsy, a rare sleep disorder that causes excessive sleepiness and frequent daytime sleep attacks, all of the top ranked drugs are drugs used in the management of insomnia. This implies that although conceptually related, the top ranked drugs are not recommended for use in narcolepsy. Likewise, in case of Netherton syndrome, a rare and severe, autosomal recessive form of ichthyosis associated with mutations in the *SPINK5* gene and currently with no known cure, the top ranked drugs in our analysis were mostly from the drug class corticosteroids. However, in practice, while topical corticosteroids may be helpful in older children, they are not usually recommended in infants as impaired barrier function in Netherton's syndrome can lead to increased cutaneous absorption resulting in complications such as pituitary adrenal axis suppression[25].

Topic Concepts as indicators for repositioning

Topic (Topic 4 – Fig. 2) shared between modafinil and biopolar disorder showed words/concepts related to neuropsychiatric or behavioral conditions (e.g., *mental Depression, major depressive disorder, attention deficit hyperactivity disorder, antidepressive agents, mental association, methylphenidate, attention, lithium, sleep*, etc.) while topic 0 shared between modafinil and narcolepsy was predominantly sleep-related (*sleep disorders, cataplexy, narcolepsy-cataplexy syndrome, sleep, REM, obstructive sleep apnea, drowsiness, hypersomnia, wakefulness, REM sleep behavior disorder,* etc.).

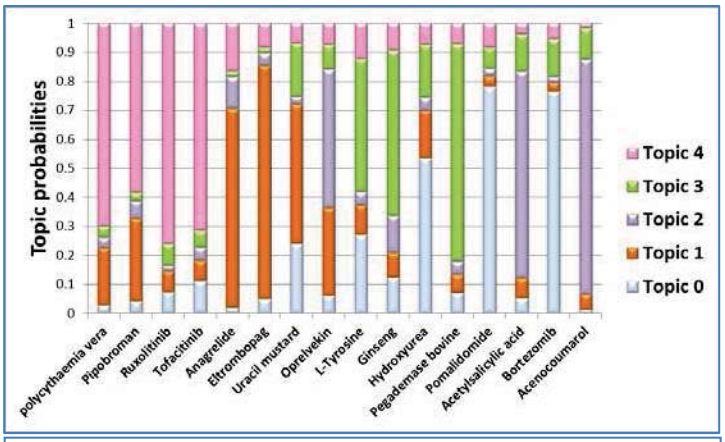


Fig. 3: Stacked bar chart showing the top ranked drugs for PV along with distribution of top five topic proportions

In case of the rare disease PV, topic 4 (Fig. 3) which was shared between the three top ranked drugs (pipobroman, ruxolitinib and tofacitinib) and PV are related to etiology of PV (e.g., *Janus kinase 2, Primary Myelofibrosis, Myeloproliferative disease, Essential Thrombocythemia, Alleles, Signal Transduction, cytokine, Janus kinase, Thrombosis, Janus kinase 1, Janus kinase 3, Interleukin-6,* etc.). Drugs ranked 4th an 5th (anagrelide and eltrombopag), interestingly are, used for thrombocytosis and thrombocytopenia and are related to PV through topic 1 (*Blood Platelets, Essential Thrombocythemia, Thrombocytopenia, Thrombocytosis, Megakaryocytes, Idiopathic Thrombocytopenic Purpura, Thrombopoiesis, Thrombosis, Thrombopoietin, Thrombus,* etc.) representing platelets and platelet-related concepts. Anagrelide reduces the platelet count and is reported to be beneficial in some patients and is recommended as second line-therapy in PV[26].

Discussion

We used topic modeling to estimate the probability distribution of topics for each of the drugs or diseases and assess the disease-drug similarity. While more extensive validation studies are required to further validate our approach, results from our preliminary validation tests and rare diseases demonstrate the utility of our approach. While the accuracy of our approach is high, the lower precision rate may be partially due to the small size of the validation sets. The novelty of our approach is several fold: first, instead of using the abstracts directly, we use mapped biomedical concepts for topic modeling which would increase the specificity and also overcome the problem of biomedical stop words to some extent; second, apart from using UMLS CUIs for topic modeling, we filter the CUIs further limiting only those belonging to relevant semantic groups; third, our approach compares disease and drug directly unlike previous approaches which focus on either drug-drug or disease-disease relationships to find drug repositioning candidates. Further, to the best of our knowledge, our study is the first to use topic modeling on MEDLINE abstracts for drug repositioning candidate discovery for rare diseases.

Some of the planned extensions for the current model relate to methodology and the data sets used. For instance, in the current study, based on the log likelihood value, we selected 200 as the topic size. However, we plan to investigate different methods and metrics for judging the optimal number of topics more systematically. While we

focused on the UMLS concepts for mapping biomedical and genomic concepts, UMLS has certain limitations especially with gene and genomic annotation representation in the UMLS Metathesaurus. We plan to supplement this by including additional resources for gene and genomic annotations (e.g., other biomedical ontologies via NCBO Annotator[27]). Since, literature related to drugs and diseases are constantly updated, the dynamic and temporal nature of the disease and drug concepts can be utilized for a more robust drug repositioning and computational pharmacovigilance systems. Although we focus on drug repositioning in this study, based on our results, the current approach can also be employed to understand the molecular basis of side-effects or suggest safer alternatives (e.g., drugs with fewer side-effects) by ranking drugs against diseases based on side-effects topics. Lastly, as a future extension, we plan to compare all of the rare disorders to approved drugs using the current approach.

References

1. Pujol A, Mosca R, Farres J, Aloy P. Unveiling the role of network and systems biology in drug discovery. Trends Pharmacol Sci. 2010;31(3):115-23.
2. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform. 2011;12(4):346-56.
3. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321(5886):263-6.
4. Hu G, Agarwal P. Human Disease-Drug Network Based on Genomic Expression Profiles. PLoS ONE. 2009;4(8):e6536.
5. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther. 2013;93(4):335-41.
6. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci U S A. 2010;107(33):14621-6.
7. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science. 2006;313(5795):1929-35.
8. Sirota M, Dudley JT, Kim J, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011;3(96):96ra77.
9. Blei DM, Ng, A.Y., Jordan, M.I. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993-1022.
10. Bisgin H, Liu Z, Kelly R, Fang H, Xu X, Tong W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. BMC bioinformatics. 2012;13 Suppl 15:S6.
11. Chasin R, Rumshisky A, Uzuner O, Szolovits P. Word sense disambiguation in the clinical domain: a comparison of knowledge-rich and knowledge-poor unsupervised methods. Journal of the American Medical Informatics Association : JAMIA. 2014;21(5):842-9.
12. Wu Y, Liu M, Zheng WJ, Zhao Z, Xu H. Ranking gene-drug relationships in biomedical literature using Latent Dirichlet Allocation. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2012:422-33.
13. Sayers E. E-utilities quick start 2008. Available from: http://www.ncbi.nlm.nih.gov/books/NBK25497/.
14. Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic acids research. 2011;39(Database issue):D1035-41.
15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium AMIA Symposium. 2001:17-21.
16. McCallum AK. MALLET: A Machine Learning for Language Toolkit 2002. Available from: http://mallet.cs.umass.edu/.
17. Kullback S. Information theory and statistics. New York: John Wiley and Sons; 1959.
18. Kiladjian JJ, Chevret S, Dosquet C, Chomienne C, Rain JD. Treatment of polycythemia vera with hydroxyurea and pipobroman: final results of a randomized trial initiated in 1980. J Clin Oncol. 2011;29(29):3907-13.
19. Passamonti F, Brusamolino E, Lazzarino M, et al. Efficacy of pipobroman in the treatment of polycythemia vera: long-term results in 163 patients. Haematologica. 2000;85(10):1011-8.
20. Verstovsek S, Passamonti F, Rambaldi A, et al. A phase 2 study of ruxolitinib, an oral JAK1 and JAK2 Inhibitor, in patients with advanced polycythemia vera who are refractory or intolerant to hydroxyurea. Cancer. 2014;120(4):513-20.
21. Harrison C, Kiladjian JJ, Al-Ali HK, et al. JAK inhibition with ruxolitinib versus best available therapy for myelofibrosis. N Engl J Med. 2012;366(9):787-98.
22. Volkers L, Kahlig KM, Verbeek NE, et al. Nav 1.1 dysfunction in genetic epilepsy with febrile seizures-plus or Dravet syndrome. Eur J Neurosci. 2011;34(8):1268-75.
23. Marini C, Scheffer IE, Nabbout R, et al. SCN1A duplications and deletions detected in Dravet syndrome: implications for molecular diagnosis. Epilepsia. 2009;50(7):1670-8.
24. Eijkelkamp N, Linley JE, Baker MD, et al. Neurological perspectives on voltage-gated sodium channels. Brain. 2012;135(Pt 9):2585-612.
25. Eichenfield LF, Tom WL, Berger TG, et al. Guidelines of care for the management of atopic dermatitis: section 2. Management and treatment of atopic dermatitis with topical therapies. J Am Acad Dermatol. 2014;71(1):116-32.
26. Finazzi G, Barbui T. Evidence and expertise in the management of polycythemia vera and essential thrombocythemia. Leukemia. 2008;22(8):1494-502.
27. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC bioinformatics. 2009;10 Suppl 9:S14.

Secure Genomic Computation through Site-Wise Encryption

Yongan Zhao XiaoFeng Wang[2] Haixu Tang[1]

School of Informatics and Computing, Indiana University, Bloomington, IN 47405

[1]: primary advisor; [2]:co-advisor

Abstract

Commercial clouds provide on-demand IT services for big-data analysis, which have become an attractive option for users who have no access to comparable infrastructure. However, utilizing these services for human genome analysis is highly risky, as human genomic data contains identifiable information of human individuals and their disease susceptibility. Therefore, currently, no computation on personal human genomic data is conducted on public clouds. To address this issue, here we present a site-wise encryption approach to encrypt whole human genome sequences, which can be subject to secure searching of genomic signatures on public clouds. We implemented this method within the Hadoop framework, and tested it on the case of searching disease markers retrieved from the *ClinVar database against patients' genomic sequences*. The secure search runs only one order of magnitude slower than the simple search without encryption, indicating our method is ready to be used for secure genomic computation on public clouds.

Background

With the advance of next-generation sequencing (NGS) technology, massive human genomic data (from whole-genome sequencing or exome sequencing) has been accumulated rapidly in laboratories and clinical settings [1]. The availability of these data accelerates the exploitation of the field of personal medicine and genome-based healthcare. To efficiently explore these datasets, research institutions and healthcare practitioners need to build infrastructure (i.e., computer systems with large memory and disk space, powerful CPUs and fast network connections) to support intensive genome computation. In addition to the high maintenance cost, human genome data comes with high liability: it contains identifiable information of human individuals (e.g., disease patients), and thus has to be protected from un-authorized access. As a result, computer servers have to be dedicated to the genome data from each specific project (or group of patients), which increases the administrative cost of the data analysis.

Commercial (also referred to as the public) clouds (such as the Amazon EC2 and the Microsoft Azure) have already been widely utilized in many fields to deliver the on-demand computation resources with pay-as-you-go pricing. Adopting an elastic computing model, commercial clouds consist of commodity machines and provide redundant data storage and high parallelism of computation capability. The cost-effectiveness of commercial clouds has been well recognized in bioinformatics for processing massive genomic data [2]. Currently, many tools are available on commercial clouds [3], for instances, reads mapping, fragment assembly and function analysis of the genomic data from microbial organisms, animals and plants. However, utilizing commercial clouds for analyzing sensitive human genomic data is hindered by the privacy concerns [4-7]. On the one hand, it is well known that even a small piece of genomic information (e.g., the genotypes over a few dozens of SNPs) can be used to infer the identity of an individual or the potential disease risk. Statistical methods [8-12] were also developed to infer the presence of a participant in a case group from the group's aggregate genomic data (e.g., allele counts on the SNP sites across the whole genome). On the other hand, commercial clouds do not offer high security assurance and tend to avoid any liability [13, 14] because in the elastic computing model, the physical servers on the cloud are shared by many users. Due to such conflict, the analysis of personal genome datasets is rarely outsourced to commercial clouds.

One can address the privacy concerns by encrypting the sensitive human genomic data before handing them over to commercial clouds. As used for storing private databases such as bank accounts on public clouds, the encrypted data provides high security assurance. There are growing interests in developing efficient protocols to support computation on encrypted data stored based on different adversary models and computing assumptions, including secure multi-party computation (SMC), oblivious RAMs, homomorphic encryption, functional encryption, property-preserving encryption, and searchable symmetric encryption (for a review see [7]).

In this paper, we present a simple cryptographic approach that supports the search of genomic signatures (e.g., one or more single nucleotide variations (SNVs) associated with a certain disease) on commercial clouds through the site-wise encryption (SWE) of whole human genomic sequences. In our adversary model, commercial clouds are reasonably assumed to be honest but curious, i.e., they are assumed to return the results for any submitted computing jobs while ensuring the correctness and integrity of results. Meanwhile, they may try to learn some statistics, such as access patterns, that can be used to infer private information from the submitted data. We also assume that the data storage can be compromised, but the computation procedure itself is protected (i.e., we do not intend to protect the computing results). Finally, we implemented our approach on the Hadoop framework; but this should not be a constraint for our approach since it can be easily extended into other platforms.

Secure Search of Genomic Signatures

We study the following genomic signature search problem: given the whole genome sequences of an individual, and a set of SNVs known to be associated with a phenotype, we want to know if the individual's genome has these SNVs (or the disease susceptibility of the individual). Our goal is to develop a secure computing protocol to solve this problem on commercial clouds, while the individual's genomic sequences are protected. More specifically, what we want to achieve are as follows: (1) public clouds do not know the content of the data, though they see how frequent each record has been queried; (2) an attacker, even when observing the queries from the user (through network eavesdropping) and obtaining encrypted SNVs (e.g., by temporarily compromising cloud storages), knows nothing about the content of the queries, SNVs and even the frequencies of the queries. We note that such a solution could be applied to several practical scenarios in genome-based medicine that require large computing resources, for examples, 1) to screen the risks of genetic diseases by searching the whole genome of an individual against the database of genetic diseases (e.g., ClinVar [15]); 2) to identify potential pharmacogenomic markers (e.g., in PharmGKB database [16]) relevant to drug dosage or selection of effective treatment; and 3) to classify patients with indistinguishable symptoms into disease subtypes based on genetic markers [17].

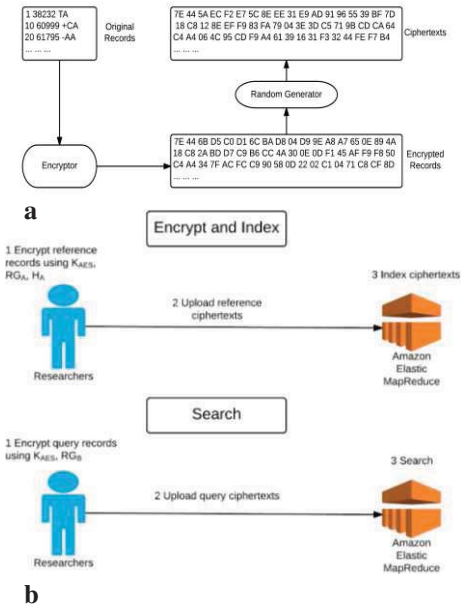


Figure 1. The cryptographic protocol for genomic signature search consisting two steps: (a) the site-wise encryption of the whole genome sequences; and (b) the secure search on public clouds.

To encrypt whole genome sequences, we first compare it against the reference human genome, and encode it as a set of variations between them, including the single nucleotide variations (substitutions or insertion/delections), long insertions/deletions, micro-inversions, and translocations. Notably, because the number of variation sites (typically several millions of them) is relatively small comparing to the whole genome sequences (3 billions of bases), this conversion effectively reduces input data size while retaining all individual genomic information, and thus is often used to compress the whole genome sequences[18]. The collection of variation sites can be encrypted effectively in a site-wise fashion: each variation site can be encrypted separately using the same encryption key. Figure 1a illustrates the encryption procedure. The data owner generates a random key, a random string generator and a hash function, and encrypts each record (including a variation, its genomic location and other information) on a local machine (see **Methods** for detailed description of the encryption protocol). The resulting ciphertexts are then uploaded to commercial clouds. When the data owner wants to search a genomic signature (e.g., one or more single nucleotide variations at specific genomic locations), he needs to encrypt the signature in the same site-wise fashion (i.e., to encrypt each variation site separately) using the same key. As a result, the same variation in the whole genome sequences will lead to the same encrypted record, although the variation itself (the type or the genomic location) cannot be inferred from it. After the encrypted genomic signature is uploaded to public clouds, a simple string matching can be conducted to determine if each variation in the genomic signature is present or not in the whole genome sequences of interest (Fig 1b). This process does not leak any information about the original records, as only encrypted data is accessed. Moreover, each variation site is unique (i.e., at a different genomic location), and thus each encrypted record is non-redundant, which prevent the inference attacks using frequency analysis on encrypted data [19].

Methods

We devised two secure search (referred to as the basic and the randomize-verify) schemes based on the following security primitives.

A symmetric key encryption function (E). A symmetric key encryption function is a family of encryption functions, which uses one same key to perform both encryption and decryption. We adopt symmetric key encryption protocol, as in our scheme, it is the data owner himself that will conduct the search on public clouds. Due to its well-designed security attributes and efficiency, AES algorithm is used in our implementation, which is a block cipher and encrypt a fixed block (128 bits) in each operation by permutation and substitution.

A cryptographic pseudorandom generator (RD). We use it to generate non-deterministic seeds and strings in our second scheme to randomize the encrypted records (Fig 1a), which further strengthen the security assurance.

A cryptographic hash function (H). A cryptographic hash function digests records and produces hash values in a collision resistant way. It is practically impossible to invert hash values to original messages. We use it in the verifi-

cation phase to determine if a pair of reference and query ciphertexts is from the same original record. SHA-256 algorithm is used in our implementation.

There are two security parameters (n and m) used in our schemes. Either n or m should be less than the length of an encrypted record, while their sum should be equal to the length of an encrypted record.

Basic scheme

In the basic scheme, the data owner first generates a random key (K_{AES}) for all records (variations) in the input genome (referred to as the reference records), and encrypts each record using E and K_{AES}. The encrypted reference records are then uploaded to public clouds after shuffling. This step needs to be done only once for a particular genome as the encrypted records can be kept on public clouds for future searches. If the data owner wants to search a genomic signature (i.e., a small set of variations), he will encrypt them using the E and K_{AES} of the corresponding genome to be searched against, and upload the encrypted records (referred to as the query records) to the cloud. Within the Map-Reduce framework (e.g., Hadoop), we use a random sampler to split the reference records into different reducers to balance the workload on every reducer. The reference records in each reducer are then indexed as a dictionary to be searched by using a binary search algorithm. When the Map-Reduce system starts to look for the exact matches between reference and query records, it automatically, in the shuffling phase, sorts query records and deliver each of them to a specific reducer, in which a binary search is initiated to determine if an exact match can be found for each query record. In the end, the match (or no match) of each query record will bel reported.

Randomize-Verify (RV) scheme

A problem of the basic scheme is that an attacker who monitors queries observes the frequencies of different queries, which could be an information leak of concern. Figure 2 illustrates an enhanced approach capable of withstanding this threat, called randomize-verify scheme. The idea is to make the ciphertexts for the different instances of the same query look different, thereby thwarting the attempt to accumulate the frequency of a specific query. Specifically, to encrypt a reference record g (Fig 2a), the data owner generates a n-bit random string (rs_R) by using RD, in addition to the key K_{AES}. He first applies E with the key K_{AES} to g, resulting in the encrypted record $E_{K_{AES}}(g)$, and then computes the cyphertext $C_R(g) = E_{K_{AES}}(g) \oplus (rs_R \cdot H_{0:m}(rs_R))$ by using the XOR operation between the encrypted record and the n-bit random string concatenated with its first m-bit hash value ($H_{0:m}(rs_R)$) computed by the hash function H, where · denote the concatenation operation on strings.

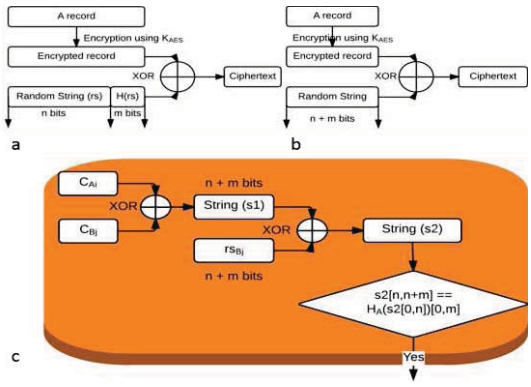


Figure 2. The randomize-verification scheme has three steps: (a) reference records encryption; (b) query records encryption; and (c) search on public clouds.

After ciphertexts are shuffled and uploaded to public clouds, they are first indexed to achieve the best performance of searching. The first 2-byte of an encrypted record ($H_{0:16}(E_{K_{AES}}(g))$) is used as a classification identifier. A random sampler is used to sample the ciphertexts to decide to which reducer a ciphertext is delivered to balance the workload of each reducer. These split databases are serialized on the local disk of each reducer so that every time when a new search task initializes, they are able to be loaded into the memory at the local machines where reducers are executed.

For the search of a genomic signature, each query record is encrypted in a similar way (Fig 2b). First, a common seed (S_Q) is generated for this specific search and a random seed (S) is generated for each query record (g'). The combination of both S_Q and S is used as the random seed to generate a $(n+m)$–bit random string ($rs_Q = RD(S_Q, S)$) for the query record g'. The ciphertext ($C_Q(g')$) is then computed by the XOR operation between the encrypted record $E_{K_{AES}}(g')$ (using E with the same K_{AES}) and the $(n+m)$–bit random string: $C_Q(g) = E_{K_{AES}}(g') \oplus rs_Q$. Finally, both $C_Q(g')$ and the corresponding S are uploaded to public clouds. The common seed can be transferred to the cloud through a secure channel (e.g., the SSH connection).

Finally, the secure search of the genomic signature can be conducted on public clouds by using a verification protocol (Fig. 2c). Every pair of reference and query ciphertext is verified (by the reducer on which the ciphertext share the same classification identifier) for representing the same variation or not by checking if $H_{0:m}(s2_{0:n}) = s2_{n:n+m}$ is true. The correctness this verification can be easily proved:

1) $s1 = C_R(g) \oplus C_Q(g) = E_{K_{AES}}(g) \oplus \left(rs_R \cdot H_{0:m}(rs_R)\right) \oplus E_{K_{AES}}(g) \oplus rs_Q = \left(rs_R \cdot H_{0:m}(rs_R)\right) \oplus rs_Q$
2) $s2 = \left(rs_R \cdot H_{0:m}(rs_R)\right) \oplus rs_Q \oplus RD(S_Q, S) = rs_R \cdot H_{0:m}(rs_R)$

so $H_{0:m}(s2_{0:n}) = s2_{n:n+m}$ if and only if $g = g'$.

Results

We implemented our schemes in Java within the Hadoop framework. It supports inputs in both plain text format (shown in Fig 1a) and variant call format (VCF). The source code of the project can be found at http://swecloud.sourceforge.net. For the testing, we used the genome sequences from a participant of Personal Genome Project (PGP) [20], from which a total of 4,005,829 variation records were retrieved. We selected 45 disease associated SNVs (i.e., query records) in ClinVar [15] to test the secure search algorithm. A full list of these SNVs and their associated diseases can be found on http://omics.informatics.indiana.edu/mg/SWECloud/, e.g., T→A substitution at 51,078,333 on chromosome 18 (ClinVar ID: rcv000021740).

In order to estimate the computation overhead of the secure search schemes, we also implemented a simple binary search algorithm to identify query records in the genome sequences in Java, and ran it on a local single CPU machine (referred to as the plain search).

Table 1 shows the performance of two secure search (basic and RV) schemes in comparison with the plain search. In the encrypt phase, the RV scheme takes longer time on reference record encryption than the basic scheme (note that this is a one-time computation), as it needs to randomize data to achieve higher security guarantee (see Discussion). The encryption for query records takes negligible time. In both of the indexing and searching phases, these two schemes show similar performances. Because the RV scheme compares each pair of query and reference ciphertexts with the same 2-byte classification identifier, the basic scheme may show slightly better performance when we have more query records. Both schemes take more time in indexing and searching comparing with the plain search. But the computation overhead is not high, at approximately one order of magnitude, which can be compensated easily with the larger computing resources available at commercial clouds.

Table 1. The performances of implemented schemes.

	Encryption (seconds)		Indexing (seconds)			Searching (seconds)		
	Reference	Query	Mapper	Reducer	Total	Mapper	Reducer	Total
Basic	37.1	0.2	59.0 (2) *	109.2 (4) *	49.0	4.6 (1)	76.5 (4) *	21.3
RV	63.7	0.2	59.4 (2)*	84.4 (4) *	46.2	4.6 (1)	75.6 (4) *	21.5
Plain	-		7.3			6.1		

(*: The numbers in parenthesis represent the numbers of parallel mapper/reducer jobs. Total running time summed over all jobs are reported.)

Discussion

The privacy risks of genome computing on public clouds are two-folded: the data stored on public clouds can be used to infer sensitive information while at each time of computation, the query data may also be identifiable. The first risks can be completely mitigated in both schemes, as the security of the encrypted records is assured by the cryptographic primitives. An adversary cannot infer useful information without knowing the encryption key even if he obtains the ciphertexts. He cannot conduct frequency analysis on the encrypted reference records either because each variation appear at most once in the genome of any single individual while the genomes of different individuals will be encrypted using different key. On the other hand, an adversary may accumulate all query ciphertexts within a period of time and then try to infer sensitive information, such as potential disease susceptibility, based on those query ciphertexts and other public information. The basic scheme may be subject to this attack. As a result, the data owners may need to change the encryption key periodically, and re-encrypt the reference records so that the adversary cannot collect the sufficient query ciphertexts to analyze the pattern. Our second scheme further decreases these risks. The reference records are double protected, by the encryption key as well as the random strings that are independent on the original records. These random strings not only protect the encrypted records, but also make ciphertexts indistinguishable. It also prevents the pattern analysis of the query ciphertexts: even if an adversary accumulates all query ciphertexts, they cannot determine if two ciphertexts in different queries are the same.

Our schemes provide a secure yet efficient way of analyzing sensitive genomic data on public clouds, based on which one can outsource not only the sensitive computation but also the sensitive genome data to clouds. The data owners (e.g., healthcare practitioners) can store encrypted genome data from all individuals (e.g., patients) on public clouds, while only keeping the encryption keys (one for each individual) locally. As such, the data owner has low liability on a large amount of sensitive data, whereas the genome data can be analyzed on clouds whenever needed.

Many schemes have been proposed for secure genome computation, most of which are based on relatively expensive cryptographic primitives (typically with >4 orders of magnitudes of computation overhead), such as secure multi-party computation (SMC) and homomorphic encryption (HE) [21, 22]. Despite their high security assurance, our

schemes based on site-wise encryption (SWE) of genomic variations are designed specifically for the secure search of human genomic data, and thus are simple and more efficient than generic schemes such as SMC and HE. Our schemes are ready to be used in real-world human genome computation, such as the genome signature search.

By adjusting n and m, we can assure the space of random strings is significantly larger than the number of encrypted records. In our implementation, we choose $n = 80$ and $m = 48$, as a person has about 4 millions of variation records in practice (thus $2^{80} \gg$ the maximum number of variation records of a person).

Acknowledgements. This work is supported by NHGRI/NIH (1R01HG007078-01) and NSF (CNS-1408874).

References

1. Ohno-Machado L. Sharing data for the public good and protecting individual privacy: informatics solutions to combine different goals. Journal of the American Medical Informatics Association. 2013;20(1):1-.
2. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11(5):207.
3. Forer L, Lipic T, Schonherr S, et al. Delivering bioinformatics MapReduce applications in the cloud. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention on; 2014: IEEE.
4. Shoenbill K, Fost N, Tachinardi U, Mendonca EA. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. Journal of the American Medical Informatics Association. 2014;21(1):171-80.
5. Lin Z, Owen AB, Altman RB. Genomic research and human subject privacy. SCIENCE-NEW YORK THEN WASHINGTON-. 2004:183-.
6. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nature Reviews Genetics. 2014;15(6):409-21.
7. Naveed M, Ayday E, Clayton EW, et al. Privacy and Security in the Genomic Era. arXiv preprint arXiv:14051891. 2014.
8. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013;339(6117):321-4.
9. Homer N, Szelinger S, Redman M, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS genetics. 2008;4(8):e1000167.
10. Craig DW, Goor RM, Wang Z, et al. Assessing and managing risk when sharing aggregate genetic variant data. Nature Reviews Genetics. 2011;12(10):730-6.
11. Sankararaman S, Obozinski G, Jordan MI, Halperin E. Genomic privacy and limits of individual detection in a pool. Nature genetics. 2009;41(9):965-7.
12. Humbert M, Ayday E, Hubaux J-P, Telenti A. Addressing the concerns of the lacks family: Quantification of kin genomic privacy. Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security; 2013: ACM.
13. Armbrust M, Fox A, Griffith R, et al. A view of cloud computing. Commun ACM. 2010;53(4):50-8.
14. Ristenpart T, Tromer E, Shacham H, Savage S. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds. Proceedings of the 16th ACM conference on Computer and communications security; Chicago, Illinois, USA. 1653687: ACM; 2009. p. 199-212.
15. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2013:gkt1113.
16. Whirl-Carrillo M, McDonagh E, Hebert J, et al. Pharmacogenomics knowledge for personalized medicine. Clinical Pharmacology & Therapeutics. 2012;92(4):414-7.
17. Bioulac‐Sage P, Rebouissou S, Thomas C, et al. Hepatocellular adenoma subtype classification using molecular markers and immunohistochemistry. Hepatology. 2007;46(3):740-8.
18. Christley S, Lu Y, Li C, Xie X. Human genomes as email attachments. Bioinformatics. 2009;25(2):274-5.
19. Zerr S, Olmedilla D, Nejdl W, Siberski W. Zerber+ r: Top-k retrieval from a confidential index. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology; 2009: ACM.
20. Church GM. The personal genome project. Molecular Systems Biology. 2005;1(1).
21. Baldi P, Baronio R, De Cristofaro E, Gasti P, Tsudik G, editors. Countering gattaca: efficient and secure testing of fully-sequenced human genomes. Proceedings of the 18th ACM conference on Computer and communications security; 2011: ACM.
22. Ayday E, Raisaro JL, Hubaux J-P, Rougemont J. Protecting and evaluating genomic privacy in medical tests and personalized medicine. Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society; 2013: ACM.