Advances in Text Mining and Visualization for Precision Medicine

Graciela Gonzalez-Hernandez†, Abeed Sarker, Karen O'Connor† and Casey Greene
Perelman School of Medicine, University of Pennsylvania
Philadelphia, Pennsylvania 19104, USA
Email: gragon@pennmedicine.upenn.edu

Hongfang Liu
Division of Biomedical Statistics and Informatics, Mayo Clinic College of Medicine
Rochester, Minnesota 55902, USA
Email: liu.hongfang@mayo.edu

According to the National Institutes of Health (NIH), precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." Although the text mining community has explored this realm for some years, the official endorsement and funding launched in 2015 with the Precision Medicine Initiative are beginning to bear fruit. This session sought to elicit participation of researchers with strong background in text mining and/or visualization who are actively collaborating with bench scientists and clinicians for the deployment of integrative approaches in precision medicine that could impact scientific discovery and advance the vision of precision medicine as a universal, accessible approach at the point of care.

Keywords: Text mining; natural language processing; precision medicine; personalized medicine; visualization; biomedicine.

1. Introduction

According to the National Institutes of Health (NIH), precision medicine is "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person." Announced in 2015, the Precision Medicine Initiative (PMI) seeks to promote research at the intersection of lifestyle, environment, and genetics to produce new knowledge for more effective ways to prolong health and treat disease [1].

Information and knowledge that could be instrumental to advances in precision medicine are buried in a vast, ever-increasing, and diverse range of data sources in structured and unstructured format: patient medical records (EMRs), standardized clinical data (such as what is required by Medicare), administrative data –from hospitals, insurance companies, and pharmacies-, patient surveys and self-reported comments from individual patients[2], the published literature, clinical trials, and research data deposited in public collections such GenBank[3] or the Gene Expression

† Work partially supported by the National Institute of Allergy And Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under grant number R01AI117011. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Omnibus (GEO) database[4], and many curated databases of interactions and pathways, to name just a few. Some recent text mining approaches related to precision medicine include automatically extracting and normalizing variant mentions in biomedical literature to reference variants in a curated database, thus allowing for analysis and novel discoveries.[5] Other relevant effort used associative text mining analysis of the free narratives of EMR to develop a system to identify previously unrecognized disease-associated factors.[6] Recent visualization advances have focused, for example, on genomic cancer data to help improve clinical decisions for precision oncology. [7–9] This session highlights original research and invited presentations on novel text mining, natural language processing (NLP), and visual analytics approaches at the intersection of lifestyle, environment, and genetics that enable further understanding of disease processes and effective treatment for individuals and cohorts that share specific characteristics.

2. Session Summary

The session includes two keynote talks by leaders in the field in biomedical data visualization and text mining, Jason Moore and Sophia Ananiadou. There are four full-length papers competitively selected for inclusion amongst the varied high-quality submissions exploring problems associated with the annotation of gene data sets, visualization of electronic health records and gene interaction to facilitate precision medicine, and concept normalization in clinical text. We selected contributions that are applicable to big genomic and text based data from multiple sources.

2.1. *Keynote: Visualization for Precision Medicine*

The first invited talk focusing on visualization is given by Jason Moore, Ph.D, Director of the Institute of Biomedical Informatics (IBI) and Senior Associate Dean for Informatics at the University of Pennsylvania's Perelman School of Medicine. Dr. Moore's work spans from artificial intelligence, data science, visualization and complex adaptive systems to systems biology, precision medicine and human genetics. Dr Moore's work relevant to this session is ample and varied. We outline here three of his most relevant papers as a quick reference:

- ViSEN, a methodology and software for visualization of statistical epistasis networks [10]. Epistasis, defined as the non-linear interaction effect among multiple genetic factors, has been recognized as a key component in understanding the underlying genetic basis of complex human diseases and phenotypic traits. ViSEN allows the analysis and visualization of two and three-way epistatic interactions. This visualized information could be very helpful to infer the underlying genetic architecture of complex diseases and to generate plausible hypotheses for further biological validations. ViSEN is freely available at https://sourceforge.net/projects/visen/.

- In PSB 2011, Dr Moore introduced a 3D visualization methodology and freely-available software package for facilitating the exploration and analysis of high-dimensional human microbiome data [11]. Powered by commercial video game development engines, the approach

provides an interactive medium in the form of a 3D heat map for exploration of microbial species and their relative abundance in different patients.

- Pioneering visualization of biological interpretation of gene expression microarray results, Dr Moore presented EVA (Exploratory Visual Analysis) [12] as a flexible combination of statistics and biological annotation to provide a visual interface for the interpretation of microarray analyses of gene expression in the most commonly occurring class of brain tumors, glioma.

Dr Moore's keynote is focused on big data processing and visualization techniques for precision medicine, and provides an insight about this rapidly emerging field. A world awash in big data presents significant computational challenges for identifying meaningful and actionable patterns. Visualization methods and technology are advancing at a rapid pace and have the potential to enable a deeper understanding of big data and derivative research results. This will require an active effort to adopt new visualization methods and to integrate them with computational analysis methods such as machine learning and natural language processing.

2.2. *Keynote: Text Mining for Precision Medicine*

The second keynote, focusing on text mining, is given by Sophia Ananiadou, PhD, director of the National Centre for Text Mining (NaCTeM) and Professor in the School of Computer Science at the University of Manchester. She has led the development of the numerous text mining tools and services currently used in NaCTeM with the aim to provide scalable text mining services: information extraction, intelligent searching, association mining, etc. She has received the IBM UIMA innovation award 3 consecutive times and is also a Daiwa award winner. Dr Ananiadou's publications relevant to this session span back at least a decade. We highlight three as a quick reference:

- In a recent publication[13], Dr Ananiadou presents a novel method that improves identification of textual uncertainty for extracted events and explores how it can be used as an additional measure of confidence for biomedical models. They use a hybrid approach that combines rule induction and machine learning with subjective logic theory to combine multiple uncertainty values extracted from different sources for the same interaction. The approach makes considerable improvements over previously published work. They evaluate their proposed system on pathways related to two different leukemia and melanoma cancer research.

- With the continuously rising need to understand the etiology of diseases as well as the demand for their informed diagnosis and personalized treatment, the curation of disease-relevant information from medical and clinical documents has become an indispensable scientific activity. Dr Anaiadou offers Argo (http://argo.nactem.ac.uk), a generic text mining workbench that can help in semi-automatic annotation of literature, including

annotation to standard terminologies, such as the UMLS. Argo's flexibility is put to the test with the semi-automatic curation of chronic obstructive pulmonary disease (COPD) phenotypes in this publication[14].

- To create, verify and maintain pathway models, curators must discover and assess knowledge distributed over vast biological literature. Dr Ananiadou explores methods for associating pathway model reactions with relevant publications[15]. The approach extracts the reactions directly from the models and then turns them into queries for three text mining-based MEDLINE literature search systems. These queries are executed, and the resulting documents are combined and ranked according to their relevance to the reactions of interest. An online demonstration of PathText 2 and the annotated corpus are available for research purposes at http://www.nactem.ac.uk/pathtext2/.

Dr Ananiadou's keynote focuses on text mining techniques to assist in the annotation and discovery of biological pathways. Pathway models are valuable resources that help us to understand the various mechanisms underpinning complex biological processes. Their curation is typically carried out through manual inspection of the scientific literature, a knowledge-intensive and laborious task. Text mining methods are used to automate model reconstruction by increasing the speed and reliability of discovery and extracting evidence from the literature. Complex information from the literature is automatically extracted and then mapped to reactions in existing pathway models. Information from the literature (events) can act as corroborative evidence of the validity of these reactions in a model or help to extend it. In addition, by contextualizing the textual evidence (extracting uncertainty, negation), we can provide additional confidence measures for linking and ranking information from the literature for model curation and ultimately better experimental design.

2.3. *Full-length Papers*

In *VisAGE: Integrating External Knowledge into Electronic Medical Record Visualization,* **Huang et al.** present a method that visualizes electronic medical records (EMRs) in a low dimensional space. Their work addresses a common issue with EMRs—that they are often fragmented and so visualization techniques often place unrelated patients close together in the visualized space. By integrating knowledge from external data sources, the system attempts to enrich EMR databases to solve this issue. This approach could aid clinicians in diagnosing and treating patients with conditions that are often misdiagnosed because they either have a collection of non-specific symptoms or are overshadowed by more prevalent conditions. The evaluations presented by the authors suggest that the method produces effective clustering of patients suffering from Parkinson's disease.

In *GeneDive: A Gene Interaction Search and Visualization Tool to Facilitate Precision Medicine,* **Previde et al.** address the problem of information overload that is faced by users of automatically mined, text-based gene interaction data by proposing a web-based tool that performs information retrieval, filtering and visualization tool. The tool, GeneDive, attempts to bring some of the best of the breed, adopting functionalities of text mining tools in the biomedical domain into a single platform. Inspired by the work of Literome [16]. GeneDive leverages Cytoscape [17], a software

package popularly used for visualization of biomolecular interactions, and DeepDive [18], a text mining tool for extracting gene interactions from literature, to provide a web-based retrieval, filtering and visualization tool for large volumes of interaction data. GeneDive offers various features and modalities that guide users through the search process to efficiently reach the information of their interest. The tool is time-efficient and it can process millions of interactions within seconds. The authors also discuss that in the future, the tool can be seamlessly extended to other interaction types such as gene-drug and gene-disease. The tool will also be made publicly available at: http://www.genedive.net.

In *Annotating Gene Sets by Mining Large Literature Collections with Protein Networks*, **Wang et al.** propose a natural language processing system that infers common functions for a gene set via the automated mining of scientific literature for relevant phrases. The system creates a heterogeneous network that connects genes with lexical concepts from the literature and combines these connections with protein interactions. The method works by performing a random walk over a heterogeneous network of phrases and genes. The authors argue that this approach presents two major advantages over previous text mining methods: (i) it integrates semantic information derived from the literature with biological information derived from experimental and interactome data, and (ii) the visualization technique reduces redundant information and visual complexity by utilizing a novel mechanism to organize functional annotations using a data structure called 'Hierarchical Concept Ontology'. The authors evaluate their method's ability to recover GO term names from the literature, applying the method to CLiXO gene sets [19], and identify a number of cancer-related terms. Evaluations of the method show substantial improvement in predicting manually curated annotations compared to a baseline text mining approach. The returned phrases remain relatively broad; however, the GO evaluation results are promising and the method takes an interesting step in the efforts to explain a gene set from literature.

In *Improving Precision in Concept Normalization*, **Boguslav et al.** propose a strategy for improving precision in medical text concept normalization by utilizing an existing high-performance biomedical concept recognition pipeline and a manually annotated corpus. The authors argue that precision is more important for health-related tasks, such as patient-centered decision support, since decisions based on false positives can be detrimental to patients' health. Although one counter-argument could be that computational system outputs are not directly used to make decisions but are vetted by human experts, and thus the role of such systems is to decrease the burden on the human agent. Thus, recall might indeed be important, but it is definitely a worthy endeavor to work toward precision gains if the loss in recall is small or can be addressed in the future, and hence the work by **Boguslav et al.** is a welcome direction. The normalization method primarily relies on a set of pre- and post-processing techniques that enable the use of a pre-existing corpus to perform the actual normalization task. The approach shows statistically significant improvements in precision over an existing baseline system for eight datasets, at the expense of recall.

3. Discussion

Text mining and visualization methods for biomedical data such as those presented in this session enable unprecedented use of data from diverse sources that can inform clinical decisions, and have come to be accepted as a necessary tool in advancing precision medicine. Visualizing such voluminous and heterogeneous data is a significant challenge, and tackling it in a way that can enable clinicians and researchers to advance precision medicine requires not only computational and logic acumen, but also creative visualization and attention to cognitive processes.

Visualization approaches and text mining techniques for information retrieval and natural language processing that are tailored to the specific needs of this domain and can handle big data play a vital role in harnessing the power of these sources to advance precision medicine research and delivery. The session aims to provide a platform for researchers to share their latest investigations in text mining and visualization and advance the vision of precision medicine as a universal, accessible approach at the point of care.

References

1. Collins FS, Varmus H. A New Initiative on Precision Medicine. *N Engl J Med*. 2015;372(9):793-795. doi:10.1056/NEJMp1500523.
2. Understanding Data Sources | Agency for Healthcare Research & Quality. https://www.ahrq.gov/professionals/quality-patient-safety/talkingquality/create/understand.html. Accessed October 6, 2017.
3. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*. 2013;41(Database issue):D36-42. doi:10.1093/nar/gks1195.
4. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207.
5. Wei C-H, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*. September 2017. doi:10.1093/bioinformatics/btx541.
6. Lin FP-Y, Pokorny A, Teng C, Epstein RJ. TEPAPA: a novel in silico feature learning pipeline for mining prognostic and associative factors from text-based electronic medical records. *Sci Rep*. 2017;7(1):6918. doi:10.1038/s41598-017-07111-0.
7. Gao J, Lindsay J, Watt S, et al. Abstract 5277: The cBioPortal for cancer genomics and its application in precision oncology. *Cancer Res*. 2016;76(14 Supplement):5277-5277. doi:10.1158/1538-7445.AM2016-5277.
8. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Sci Signal*. 2013;6(269):pl1. doi:10.1126/SCISIGNAL.2004088.
9. Klonowska K, Czubak K, Wojciechowska M, et al. Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget*. 2016;7(1):176-192. doi:10.18632/oncotarget.6128.
10. Hu T, Chen Y, Kiralis JW, Moore JH. ViSEN: methodology and software for visualization of statistical epistasis networks. *Genet Epidemiol*. 2013;37(3):283-285. doi:10.1002/gepi.21718.
11. Moore JH, Lari RCS, Hill D, Hibberd PL, Madan JC. Human microbiome visualization using 3D technology. *Pac Symp Biocomput*. 2011:154-164. http://www.ncbi.nlm.nih.gov/pubmed/21121043. Accessed October 6, 2017.
12. Reif DM, Israel MA, Moore JH. Exploratory Visual Analysis of statistical results from microarray experiments comparing high and low grade glioma. *Cancer Inform*. 2007;5:19-24. http://www.ncbi.nlm.nih.gov/pubmed/19390666. Accessed October 6, 2017.
13. Zerva C, Batista-Navarro R, Day P, Ananiadou S. Using uncertainty to link and rank evidence from biomedical literature for model curation. *Bioinformatics*. July 2017. doi:10.1093/bioinformatics/btx466.
14. Batista-Navarro R, Carter J, Ananiadou S. Argo: enabling the development of bespoke workflows and

services for disease annotation. *Database (Oxford)*. 2016;2016. doi:10.1093/database/baw066.
15. Miwa M, Ohta T, Rak R, et al. A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. *Bioinformatics*. 2013;29(13):i44-52. doi:10.1093/bioinformatics/btt227.
16. Poon H, Quirk C, DeZiel C, Heckerman D. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics*. 2014;30(19):2840-2842. doi:10.1093/bioinformatics/btu383.
17. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi:10.1101/gr.1239303.
18. Mallory EK, Zhang C, Ré C, Altman RB. Large-scale extraction of gene interactions from full-text literature using DeepDive. *Bioinformatics*. 2016;32(1):106-113. doi:10.1093/bioinformatics/btv476.
19. Kramer M, Dutkowski J, Yu M, Bafna V, Ideker T. Inferring gene ontologies from pairwise similarity data. *Bioinformatics*. 2014;30(12):i34-42. doi:10.1093/bioinformatics/btu282.