PATTERNS IN BIOMEDICAL DATA-HOW DO WE FIND THEM?

ANNA O. BASILE
The Pennsylvania State University, Department of Biochemistry and Molecular Biology
328 Innovation Blvd Ste 210
State College, PA 16803
azo121@psu.edu

ANURAG VERMA
Geisinger Health System
The Pennsylvania State University, Huck Institutes of the Life Sciences
328 Innovation Blvd Ste 210
State College, PA 16803
averma@geisinger.edu

MARTA BYRSKA-BISHOP
Geisinger Health System
328 Innovation Blvd Ste 210
State College, PA 16803
mbyrskabishop@geisinger.edu

SARAH A. PENDERGRASS
Geisinger Health System, Biomedical and Translational Informatics
122 Weis Center for Research
Danville, PA 17822
spendergrass@geisinger.edu

CHRISTIAN DARABOS
Dartmouth College, Research Computing Services
HB 6129
Hanover, NH 03755
christian.darabos@dartmouth.edu

H. LESTER KIRCHNER
Geisinger Health System, Biomedical and Translational Informatics
100 N. Academy Ave
Danville, PA 17822-4400
hlkirchner@geisinger.edu

Given the exponential growth of biomedical data, researchers are faced with numerous challenges in extracting and interpreting information from these large, high-dimensional, incomplete, and often noisy data. To facilitate addressing this growing concern, the "Patterns in Biomedical Data-How do we find them?" session of the 2017 Pacific Symposium on Biocomputing (PSB) is devoted to exploring pattern recognition using data-driven approaches for biomedical and precision medicine applications. The papers selected for this session focus on novel machine learning techniques as well as applications of established methods to heterogeneous data. We also feature manuscripts aimed at addressing the current challenges associated with the analysis of biomedical data.

1. Introduction

With great technological advances and numerous ‘big data’ initiatives targeted at generating and acquiring large amounts of biomedical information, there has been an astonishing growth in the volume of data in recent years [1]. Considering sequencing data alone, the size of data has approximately doubled every six months in the last decade [2]. Continuing at this rate, we can expect to reach a zettabyte of sequencing data generated per year by 2025 [2].

Thus, the age of big data is upon us, and with its arrival comes the potential to revolutionize many aspects of our lives. Decisions previously made using carefully constructed, simulated models of reality can now be made using measured data. While the term ‘big data’ is not well defined, it will be used herein to describe a situation where the amount of information far exceeds that which has been previously available [3]. Big data analyses impact many areas of society, culture, and research. To combat crime, law enforcement officials are employing seismology-like models to predict areas of high crime, and intervene to prevent them from occurring [3]. With large scale surveys, such as the Two Micron All-Sky Survey, which contains a petabyte of data, astronomers can now focus their efforts illuminating structures and exploring potential connections and hypotheses [4]. In the area of public health and precision medicine, large-scale efforts have been made to create datasets aimed at elucidating the genetic underpinnings of various traits as a means of disease prevention and development of effective treatment. For example, the Precision Medicine Initiative Cohort Program announced by President Obama plans to enroll one million participants spanning a multitude of age and race groups within the US [5]. Other large-scale genome projects include the UK 100,000 Genomes Project [6], and the Geisinger MyCode Community Health Initiative which unites Geisinger Health System and Regeneron Genetics Center in a collaboration aimed at bio-banking and whole-exome sequencing more than 200,000 patients [7]. Likewise, public datasets, such as The Cancer Genome Atlas (TCGA), which provides molecular characterization of cancer genomes, continue to provide a wealth of data to researchers with the hope of one day improving clinical patient care.

While these potentials are truly revolutionary, there are a number of challenges that can impede the promises of big data and make it difficult to extract the true value of this information. The sheer volume of available data and the rate at which it is being generated is overwhelming the majority of industries, many of which do not yet have the proper management, storage and analytical means of assessing this information [8]. Additionally, while small sample sizes are often prohibitive in research, the large sample sizes provided by big data initiatives may not be a panacea. Large sample sizes may be of little value if they are not representative of the population being assessed, are missing information (especially if missingness is nonrandom or important data is completely missing), or contain sampling biases [9]. Machine-learning approaches in this data-driven space will require an integration of different generated data types. In a biomedical setting, this may include clinical measurements, drug usage data, mRNA expression levels, and environmental exposures. These informatics methods must also be robust to incompleteness and variable sparsity, as well as heterogeneity which can present mixtures of categorical and numerical data. Further considerations that will need to be made include scalability and dealing with a feature space that far exceeds the number of samples.

The collection of papers presented in this session demonstrates a diversity of data-driven, pattern recognition approaches and challenges within the biomedical and precision health setting. These manuscripts span a wide range of categories from applications of well-studied informatics methods to novel pattern recognition techniques as well as approaches of overcoming big data challenges.

2. Session Contributions:

2.1 Machine Learning and Deep Learning Approaches

Machine learning and deep learning have recently received a great deal of attention due to their potentially transformative applications to big data. Machine learning refers to a class of algorithms that can learn from and also make predictions on data [10], while deep learning describes a branch of machine learning that models data using multiple levels of representation and abstraction. These methods do not require explicit rules as they rely on the data, and generally speaking, the more data, the better the outcome of these techniques. While the use of data-driven approaches is not new, this is an expanding area of biomedical research that is gaining momentum due to algorithmic sophistication, computational advancement, and the growth in volume and variety of available data.

Shameer et al. describe a data-driven feature selection and machine learning approach to predict hospital readmission in heart failure (HF) patients from electronic health records (EHR) in "*Predictive Modeling of Hospital Readmission Rates using Electronic Medical Record-wide Machine Learning: A Cased-Study Using Mount Sinai Health Cohort".* Several data domains were extracted from the EHR including diagnoses, medications, laboratory measurements, procedures, and vitals. Separate models were generated from the data domains using the Naïve Bayes algorithm and then combined. Feature selection was performed using a correlation-based method. Their approach was contrasted to using logistic regression, and it performed well over all existing predictive models in HF.

In the manuscript "*Missing data imputation in the electronic health record using deeply learned autoencoders*" **Beaulieu-Jones** et al. tackle the important issue of dealing with missing data, commonly encountered in the context of EHR. Specifically, the authors use the Pooled Resource Open-Access Amyotrophic Lateral Sclerosis (ALS) Clinical Trial Database (PRO-ACT) to evaluate missing data imputation performance of a machine learning approach, namely deeply learned autoencoders, and compare it to the performance of several established imputation strategies, such as mean, median, K-nearest neighbors, or Singular Value Decomposition (SVD). They show that autoencoders outperform other methods in imputation of data missing completely at random (MCAR), as well as data missing not at random (MNAR). Furthermore, they used data imputed by different methods to predict ALS progression and identify the most important predictors of ALS.

One of the challenges associated with applying machine learning approaches to biological problems is the interpretation of the models that arise from them. In the manuscript titled "*DG-Dashboard: Visualizing and Understanding Genomic Sequences Using Deep Neural Networks*", **Lanchantin** et al. present a visualization toolkit called the Deep Genomic Dashboard (DG-Dashboard), which facilitates interpretation of deep neural network models in the context of predicting transcription factor binding sites (TFBS) along genomic DNA. In particular, DG-Dashboard offers three strategies: saliency maps, temporal output scores, and class optimizations, which enable visualization of nucleotide importance within a particular motif, critical positions along a DNA sequence, as well as class-specific motif patterns for a particular TF based on predictions obtained from convolutional neural networks (CNNs), recurrent neural networks (RNNs), as well as convolutional-recurrent neural networks (CNN-RNNs). In addition to facilitating interpretation of the three deep neural network architectures, Lanchantin et al. demonstrate that CNN-RNNs outperform CNN and RNN in classification of TFBSs.

2.2 Pattern recognition applications in EHR, Medical Imaging, and Mobile Health data

Applications of machine learning approaches are widespread in the biomedical sector. EHRs, biomedical images, and mobile health apps are just a few of the many sources researchers are mining to advance human health. Data-driven approaches can leverage the wealth of information in these sources and extract meaningful knowledge which can then be utilized to study disease progression and symptom patterns, classify patient subgroups, and inform clinical practice and decision-making.

One such application is digital image analysis that was implemented to classify the bone cancer in “*Large Scale Image Segmentation and Classification for Viable and non-viable Tumor Identification in Osteosarcoma*”. **Arunachalam** et al. demonstrate a high-throughput approach to classify the tumor region from images of Hematoxylin and eosin (H&E) stain slides from bone cancer patients. They proposed a multi-tier approach where they used pixel and object based approach to color and classify different histopathological regions of cancer cells in the digital stain images. Further, they used a combination of multiple clustering algorithms to define viable and non-viable tumors.

In “*Development and Performance of Text-Mining Algorithms to Extract Socioeconomics Status from DE-identified Electronic Health Records*”, **Hollister** et al. describe a data mining approach, where they developed an algorithm to define a phenotype status from variety of structured and unstructured free text in EHR. In order to investigate socioeconomics status (SES) they developed seven different algorithms predictive of SES like Education, Occupation, Insurance Status, Retirement, Medicaid, and Homelessness. Their work addresses an important question associated with health outcomes and the socioeconomic status extracted from various semantic categories. They provide performance metric of seven algorithms, but also highlight many shortcoming and challenges that potentially affect phenotype algorithm development in current EHR systems.

In “*Methods for Clustering Time Series Data Acquired from Mobile Health Apps*”, **Tignor** and colleagues present a method to cluster individuals with asthma using data collected from a mobile health app. The data represent a time series of daily asthma symptoms which exhibit non-ignorable missingness. Their work focuses on developing a novel probabilistic imputation method, and combined with a consensus clustering algorithm, is used to identify distinct symptom patterns. Variations on the algorithm implementation are devised and compared.

Studying the heterogeneous patterns of disease manifestation and progression is important for the clinical treatment and management of a condition. In “*Learning Attributes of Disease Progression from Trajectories of Sparse Lab Values*”, **Agarwal** et al. use the Functional Clustering Model (FCM) to cluster sparse clinical lab measures from patients with Chronic Kidney Disease (CKD) from the Stanford Health Care (SHC) system. The authors hypothesize that using data-driven approaches on trajectories of sparse lab values can create clinically meaningful clusters that highlight alternate disease progression patterns in CKD. Irregularity and sparsity in longitudinal EHR data creates high variance in trajectory estimates and often leads to unstable clusters. The FCM approach addresses this challenge by treating curve coefficients as random effects, and then projecting the curve into a subspace where the cluster centers now represent the probability of cluster membership. Using this approach, the authors cluster creatinine trajectories of CKD patients to create two patient groupings which feature distinct clinical attributes.

2.3 Public Data Mining

The extraction and identification of higher level relationships from high-throughput data and data repositories is an important area of research. For example, with the ever increasing amount of study information existing within PubMed, it is a challenge to integrate that much information to gain higher level insights over trends that have been found for genes and diseases. The information gained from effectively integrating comprehensive data together in novel ways could ultimately result in the “sum being greater than the parts”, providing new insights for further research and discovery.

In “*A new relevance estimator for compilation and visualization of disease patterns and potential drug targets*”**, von Korff** et al. describe a tool, the Disease Relevance Miner (DDRelevanceMiner), which was developed using the concept of second order co-occurrence which takes advantage of calculating the similarity between two words that do not co-occur frequently, but co-occur with the same neighboring word. The authors used the basis of this approach but with the advancement of a relevance estimator. Using the DDRelevance Miner, the authors used HUGO gene identifiers, and then linked them to PubMed in order to extract relevant records for each gene, where each publication record in turn was searched with disease MeSH terms. Linking together these data along with a metric of relevance, provided detailed disease-gene and disease-disease associations which could be further explored. This includes the identification of gene drug targets that had indications of being highly specific to single diseases.

Wilson et al. evaluate the performance of four community detection algorithms to automatically determine groups of genes from protein-protein interaction networks using experimental data in "*Discovery of Functional and Disease Pathways by Community Detection in Protein-Protein Interaction Networks*". To date, biological pathway information has been based on experimentally gained understanding. The various pathway repositories that exist are incredibly important resources, a testament to how much has been learned of the underlying structure of biology. These resources contribute to a greater understanding of gene expression and genetic association results, as well as identification of genetic interaction candidates. High throughput computational approaches could help fast track the evaluation of new potential pathways. Determining communities of biological networks could shed new light on groupings of genes with common biological functions or features. With the reliance of many analyses based on gene and pathway information, such as the Gene Set Enrichment Analysis (GSEA) [11], Pathway Analysis by Randomization Incorporating Structure (PARIS) [12], and other tools like Biofilter [13], further identification of pathways could support new hypothesis generation for experimental validation. In the manuscript by Wilson et al., several possible community detection methods were tested using a STRING protein-protein interaction network [14]. Communities obtained were then compared to curated biological pathways, over multiple metrics. Both known pathways were re-identified and possibly novel pathways were identified, the authors carefully characterized other features of these networks as well, highlighting the utility of community detection methods in identifying new pathways for further study.

References:

1. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, et al. The NIH Big Data to Knowledge (BD2K) initiative. J. Am. Med. Inform. Assoc. 2015;22:1114–1114.

2. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13:e1002195.

3. Hvistendahl M. Can "predictive policing" prevent crime before it happens? Sci. Mag. [Internet]. 2016; Available from: http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens

4. Zhang Y, Zhao Y. Astronomy in the Big Data Era. Data Sci. J. 2015;14:11.

5. PMI in the News [Internet]. Natl. Inst. Health NIH. 2015 [cited 2016 Sep 23]. Available from: https://www.nih.gov/node/19706/draft

6. Rabes T, 1 ratanaAug, 2014, Pm 12:30. U.K.'s 100,000 Genomes Project gets £300 million to finish the job by 2017 [Internet]. Sci. AAAS. 2014 [cited 2016 Sep 23]. Available from: http://www.sciencemag.org/news/2014/08/uks-100000-genomes-project-gets-300-million-finish-job-2017

7. Carey DJ, Fetterolf SN, Davis FD, Faucett WA, Kirchner HL, Mirshahi U, et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. 2016;18:906–13.

8. Kaplan RM, Chambers DA, Glasgow RE. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. Clin. Transl. Sci. 2014;7:342–6.

9. DeRouen TA. Promises and Pitfalls in the Use of "Big Data" for Clinical Research. J. Dent. Res. 2015;94:107S – 109S.

10. Vadrevu S. Understanding the Promise and Pitfalls of Machine Learning [Internet]. Data Inf. 2015 [cited 2016 Sep 30]. Available from: http://data-informed.com/understanding-promise-pitfalls-machine-learning/

11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 2005;102:15545–50.

12. Butkiewicz M, Cooke Bailey JN, Frase A, Dudek S, Yaspan BL, Ritchie MD, et al. Pathway analysis by randomization incorporating structure-PARIS: an update. Bioinforma. Oxf. Engl. 2016;32:2361–3.

13. Pendergrass SA, Frase A, Wallace J, Wolfe D, Katiyar N, Moore C, et al. Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. BioData Min. 2013;6:25.

14. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. 2013;41:D808–15.