Session Introduction
Pattern Recognition in Biomedical Data: Challenges in putting big data to work

Shefali Setia Verma
University of Pennsylvania
Philadelphia, PA 19104

Anurag Verma
University of Pennsylvania
Philadelphia, PA 19104

Dokyoon Kim
Geisinger
100 North Academy Avenue
Danville, PA 17822

Christian Darabos
Research Computing Services, Dartmouth College,
HB 6129
Hanover, NH 03755

Introduction

Technological advances are leading to an exponential increase in the size of biomedical data. Demand is high for novel computational techniques that can cope with these large datasets and have the potential to support translational research. Methods to analyze biomedical data in order to handle its complexities require sophisticated algorithms for pattern recognition and to handle complexities such as sparesenss and noisiness in these datasets. The availability of high throughput techniques in generating highly resourceful multi-omic biomedical data (genomic, transcriptomic and epigenomic to name a few) gave rise to a whole new set of challenges in identifying patterns. Modern statistical, machine learning, and even artificial intelligence (AI) methods can be used to integrate multiple resources to understand complex phenotypic traits. However, most of these methods pose multiple challenges either in fitting models or in analyzing the resulting models, whether using multiple species or multi-omic datasets for the

same species. This session focuses on innovative ways to address the challenges arising from the quality and quantity of data and also integrating biomedical data from various sources to identify patterns in biomedical datasets[1–3].

While cloud computing aids in analysis performance by improving computing time and storage, it is limited to the software package and there is considerable room for improvement in the cloud-based big-data analysis. Our session also aims at discussing the optimization of tool development for large scale datasets and challenges that are associated with the computational cost as well as resources for pattern recognition. Manuscripts listed in this session can be classified into following 4 categories:

1. ***Identifying patterns in EHR data***:
   Electronic Health Records (EHRs) is a collection of longitudinal health information from an individual's point of care. It includes diagnosis, procedure, laboratory measurement, medication, imaging, and clinical note. Many retrospective case-control studies have already demonstrated meaningful use of EHR data and its potential to improve understanding of disease risk and prevalence in the general populatio[4–7]. However, the data within EHR has not been utilized to its full extent due to several challenges, such as missing data, institutional biases in coding practice, and high throughput electronic phenotyping.

   In the manuscript titled "*Learning Contextual Hierarchical Structure of Medical Concepts to Clarify Phenotypes*", *Beaulieu-Jones et al* present an innovative application of Pointcaré embeddings to model data-driven hierarchy of ICD-9 diagnosis codes. The Pointcaré embeddings approach uses hyperbolic space to learn the embedding from a vector of nodes in a network graph as opposed to traditional Euclidean space-based methods such as Word2Vec[8]or GloVe[9] Since it is shown that the hyperbolic space is more appropriate for hierarchical information[10], so its application of ICD-9 codes shows potential in improving phenotype definitions while keeping the global structure and hierarchy of ICD-9 codes.

   Similarly, as the new methods are showing improvement in electronic phenotyping in EHR data, it is also important to identify patient cohort for a disease more accurately. In manuscript titled "*The Effectiveness of Multitask Learning for Phenotyping with*

Electronic Health Records Data", *Ding et al* investigated the effectiveness of a supervised approach called Multitask Learning (MTL) to define phenotypes using EHR data. Authors demonstrated that MTL approach performed better for complex phenotype definition whereas traditional supervised approaches such as linear models can be preferable for simple phenotype definitions.

Integrating EHR data from various health providers across the country has great potential to predict disease risk across the large population. However, there are various disparities across different health providers such as clinical care bias, population differences, ethical, and privacy policies. In the manuscript "*ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites*", *Duan et al* propose an algorithmic approach to integrate EHR data from multiple health providers in an efficient way, and preserving privacy. They propose a use of a common data model developed by Observational Health Data Sciences and Informatics (ODSHI) and further perform statistical analysis in a distributed manner across multiple sites. Authors address a key issue of data sharing using ODAL by performing large-scale association analysis without explicitly sharing of sensitive data.

2. ***Machine/Deep Learning approaches:***

The current explosion of biomedical big data, including imaging, genomic, and EHR, provide a great opportunity to improve understanding of the genetic architecture of complex diseases and ultimately to improve health care. With the explosion of the biomedical big data, machine learning and deep learning techniques are becoming an integral component of evaluating biomedical data. In particular, deep learning has been extensively used in the field of biomedical informatics, such as healthcare and genomic data analyses as well as text mining.

In the context of healthcare data analysis, the accurate detection of premature ventricular contractions (PVC) in patients is an important task in cardiac care for some patients. *Gordon et al* developed a novel PVC detection algorithm based around a convolutional autoencoder to address the weaknesses, such as the need to use difficult to extract morphological features, domain-specific features, or large number of estimated parameters, and validated their method using the MIT-BIH arrhythmia

database. Although many deep learning methods have been shown with great successes in biomedical informatics, the "black-box" nature of deep learning and the high-reliability requirement of biomedical applications have created new challenges regarding the existence of confounding factors. In the manuscript titled "Removing Confounding Factors Associated Weights in Deep Neural Networks Improves the Prediction Accuracy for Healthcare Applications", *Wang et al* present an efficient method that can remove the influences of confounding factors, such as age or gender, to improve the across-cohort prediction accuracy of deep neural networks.

Deep learning is also applied to many genomic data analyses. Protein domain boundary prediction is usually an early step to understand protein function and structure. Most of the current computational domain boundary prediction methods suffer from low accuracy and limitation in handling multi-domain types, or even cannot be applied on certain targets, such as proteins with the discontinuous domain. *Jiang et al* developed an *ab-initio* protein domain predictor using a stacked bidirectional Long Short-Term Memory Units (LSTM) model in deep learning. Additionally, a deep residual network (deep ResNet) is a type of specialized neural network that helps to handle more sophisticated deep learning tasks and models. *Liu et al* describe the use of a deep ResNet-based model that fuses flanking DNA sequence information with additional SNP annotation information for identifying functional noncoding SNPs in trait-associated regions. As another interesting study, steganography serves to conceal the existence and content of messages in the media using various techniques. Recent advances in next-generation sequencing technologies have facilitated the use of deoxyribonucleic acid (DNA) as a novel covert channel in steganography. *Bae et al* propose a general sequence learning-based DNA steganalysis framework using deep recurrent neural networks (RNNs). The proposed approach learns the intrinsic distribution of coding and non-coding sequences and detects hidden messages by exploiting distribution variations after hiding these messages.

In addition to many applications, deep learning technique is widely used in text mining. Phylogeography research involving virus spread and tree reconstruction relies on accurate geographic locations of infected hosts. Insufficient level of geographic information in nucleotide sequence repositories such as GenBank motivates the use of

natural language processing methods for extracting geographic location names (toponyms) in the scientific article associated with the sequence and disambiguating the locations to their coordinates. *Magge et al* present an extensive study of multiple recurrent neural network architectures for the task of extracting geographic locations and their effective contribution to the disambiguation task using population heuristics. Additionally, in the manuscript titled "Automatic Human-like Mining and Constructing Reliable Genetic Association Database with Deep Reinforcement Learning", *Wang et al* aim to improve the reliability of biomedical text-mining by training the system to directly simulate the human behaviors, such as querying the PubMed, selecting articles from queried results, and reading selected articles for knowledge. They take advantage of the efficiency of biomedical text-mining, the flexibility of deep reinforcement learning, and the massive amount of knowledge collected in UMLS into an integrative artificial intelligent reader that can automatically identify the authentic articles and effectively acquire the knowledge conveyed in the articles.

Although classification has been extensively studied over the past decades, there remain understudied problems when the training data violate the main statistical assumptions relied upon for accurate learning and model characterization. This particularly holds true in the open world setting where observations of a phenomenon generally guarantee its presence, but the absence of such evidence cannot be interpreted as the evidence of its absence. Learning from such data is often referred to as positive-unlabeled learning, a form of semi-supervised learning where all labeled data belong to one (say, positive) class. To improve the best practices in the field, *Ramola et al* study the quality of estimated performance accuracy in positive-unlabeled learning in the biomedical domain.

3. *Identifying patterns in omic data sets:*

Complex traits are often heterogeneous in nature, which means that they are likely not only explained by one data type (for example genomic variations). Thus, integrative methods in combining data from various sources (on same or different samples) is demanding. *Graim et al* present a new method for integrating multiple data types to predict cancer-drug sensitivity. The proposed method PLATYPUS (Progressive LAbel Training bY Predicting Unlabeled Samples) combines prior knowledge with raw input data to make predictions in testing dataset. This method when compared to

ensemble approach on using single dataset yields better prediction even in samples where missingness is observed. *Marty et al* represent an integrative approach for utilizing exome and transcriptome to study the highly heterogeneous Killer Immunoglobulin-like receptor (KIR) region that is known to be associated with cancer phenotypes. Lastly, *Pyman et al* use deep learning methods to classify 26 types of cancer cells from normal tissue cell by analyzing microRNA dataset.

Understanding gene function is an important aspect of interpretation of findings. Rapid advancements have been made in sequencing microbial genome. *Li et al* present a Bayesian approach to analyze transposon mutagenesis with next generation sequencing (TnSeq) data. *Anand et al* represent a method to link non-coding variants to gene functions by using CHIP-Seq data for interpreting association study signals.

Publicly available open source large datasets also provide unique opportunities for pattern recognition. Leveraging these resources are highly important. *Tsui et al* utilized datasets from Sequence Read Archive (SRA) and designed a pipeline to extract allele counts from variety of datasets, such as RNA seq, whole exome sequencing and whole genome sequencing.

4. ***Computational challenges:***

The data-intensive nature of the computational problem in the field of biomedical informatics also warrants the development of software approaches to efficiently use the existing institutional computer infrastructure as well as cloud computing. Additionally, the tools and workflows are changing at a rapid pace as new data types are being generated from new techniques in biology such as sequencing, gene expression data, among others. This raises two key issues: assessment of new software workflows and their reproducibility. There is community effort like Dialogue for Reverse Engineering Assessments and Methods (DREAM) Challenges to compare and benchmark new tools and workflows. In the manuscript "*A Workflow-based Approach to Benchmark Challenges Enhances Reusability, and Reproducibility*", *Srivastava et al* present an approach to improve the reproducibility and interpretability associated with bioinformatics benchmark challenges. To achieve this, the authors used the WINGS system as the model and modified it to allow each step of the submitted algorithms to be analysed.

References

1. Bourne, P. E.; Bonazzi, V.; Dunn, M.; Green, E. D.; Guyer, M.; Komatsoulis, G.; Larkin, J.; Russell, B. The NIH Big Data to Knowledge (BD2K) initiative. *J Am Med Inform Assoc* **2015**, *22*, 1114, doi:10.1093/jamia/ocv136.
2. Ritchie, M. D.; Holzinger, E. R.; Li, R.; Pendergrass, S. A.; Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97, doi:10.1038/nrg3868.
3. Pasaniuc, B.; Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **2017**, *18*, 117–127, doi:10.1038/nrg.2016.142.
4. Verma, A.; Verma, S. S.; Pendergrass, S. A.; Crawford, D. C.; Crosslin, D. R.; Kuivaniemi, H.; Bush, W. S.; Bradford, Y.; Kullo, I.; Bielinski, S. J.; Li, R.; Denny, J. C.; Peissig, P.; Hebbring, S.; De Andrade, M.; Ritchie, M. D.; Tromp, G. eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants. *BMC Medical Genomics* **2016**, *9*, 32, doi:10.1186/s12920-016-0191-8.
5. Verma, S. S.; Lucas, A. M.; Lavage, D. R.; Leader, J. B.; Metpally, R.; Krishnamurthy, S.; Dewey, F.; Borecki, I.; Lopez, A.; Overton, J.; Penn, J.; Reid, J.; Pendergrass, S. A.; Breitwieser, G.; Ritchie, M. D. IDENTIFYING GENETIC ASSOCIATIONS WITH VARIABILITY IN METABOLIC HEALTH AND BLOOD COUNT LABORATORY VALUES: DIVING INTO THE QUANTITATIVE TRAITS BY LEVERAGING LONGITUDINAL DATA FROM AN EHR. *Pac Symp Biocomput* **2016**, *22*, 533–544.
6. Hoffmann, T. J.; Ehret, G. B.; Nandakumar, P.; Ranatunga, D.; Schaefer, C.; Kwok, P.-Y.; Iribarren, C.; Chakravarti, A.; Risch, N. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. *Nat Genet* **2017**, *49*, 54–64, doi:10.1038/ng.3715.
7. Singh, A.; Nadkarni, G.; Gottesman, O.; Ellis, S. B.; Bottinger, E. P.; Guttag, J. V. Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration. *J Biomed Inform* **2015**, *53*, 220–228, doi:10.1016/j.jbi.2014.11.005.
8. Mikolov, T. Efficient Estimation of Word Representations in Vector Space., doi:arXiv:1301.3781.
9. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543.
10. Krioukov, D.; Papadopoulos, F.; Kitsak, M.; Vahdat, A.; Boguñá, M. Hyperbolic geometry of complex networks. *Physical Review E* **2010**, *82*, doi:10.1103/PhysRevE.82.036106.