Session Introduction: Challenges of Pattern Recognition in Biomedical Data

Shefali Setia Verma
Geisinger Health System
The Huck Institute of the Life Sciences, The Pennsylvania State University,
328 Innovation Blvd Ste 210
State College, PA 16803

Anurag Verma
Geisinger Health System
The Huck Institute of the Life Sciences, The Pennsylvania State University
328 Innovation Blvd Ste 210
State College, PA 16803

Anna Okula Basile
Department of Biochemistry and Molecular Biology, The Pennsylvania State University
328 Innovation Blvd Ste 210
State College, PA 16803

Marta-Byrska Bishop
Geisinger Health System
328 Innovation Blvd Ste 210
State College, PA 16803

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

The analysis of large biomedical data often presents with various challenges related to not just the size of the data, but also to data quality issues such as heterogeneity, multidimensionality, noisiness, and incompleteness of the data. The data-intensive nature of computational genomics problems in biomedical informatics warrants the development and use of massive computer infrastructure and advanced software tools and platforms, including but not limited to the use of cloud computing. Our session aims to address these challenges in handling big data for designing a study, performing analysis, and interpreting outcomes of these analyses. These challenges have been prevalent in many studies including those which focus on the identification of novel genetic variant-phenotype associations using data from sources like Electronic Health Records (EHRs) or multi-omic data. One of the biggest challenges to focus on is the imperfect nature of the biomedical data where a lot of noise and sparseness is observed. In our session, we will present research articles that can help in identifying innovative ways to recognize and overcome newly arising challenges associated with pattern recognition in biomedical data.

1. Introduction:

Machine learning methods are designed to identify regularities in datasets and then use the identified patterns in a subset of the data to make predictions for the rest of the data. Supervised and unsupervised machine learning methods for pattern recognition have been widely applied in many fields such as image and speech recognition, medical diagnosis, business analytics, finance, as well as in social media, movie recommendations (Netflix), retails, to name a few[2]. With technological advancements, biomedical data is increasing exponentially in size, and there is a high demand to apply these techniques to understand the etiologies of complex diseases[3]. To achieve this goal, it is important to address the challenges of big data analytics and develop optimized methods for pattern recognition that can handle complexities of biomedical data.

The biomedical field, in the current era of precision medicine, is recognized for the interest of researchers in elucidating the genetic architecture of human traits/diseases to improve clinical care. Some of the publicly available 'Big Data' datasets include but are not limited to The 1000 Genomes Project, The Cancer Genome Atlas (TCGA), UK Biobank, Encyclopedia of DNA Elements (ENCODE), Gene Expression Omnibus (GEO), the Library of Integrated Network-based Cellular Signatures (LINCS), the database of Genotypes and Phenotypes (dbGaP), and many other[4–7]. These resources consist of metadata from association analyses, variant information from commercial genotyping chips, whole exome and genome sequencing data, phenotype information, structural variation, gene expression, and among others. Challenges for identifying patterns arise in one data type and increases more in attempts to integrate multiple aforementioned data types/omics[8]. This expanding knowledge is both a blessing and a curse for identifying patterns. Traditional methods of analyzing biomedical data obtained from various high throughput sources are inadequate to handle the ever-increasing wealth of knowledge that is gathered about genotype and phenotype. In this session, we will address the challenges arising from attempts to integrate biomedical data from various sources (including, but not limited to, one or across more species, use of raw data, or summary level statistics) and identify patterns from these multi-omic datasets[9].

The data-intensive nature of computational genetics problem sets in the biomedical informatics field warrants the development and use of vast computer infrastructure and advanced software tools and platforms. Many existing technologies, e.g., Hadoop, Spark, MongoDB, Neo4j, make storage and analytics of large-scale datasets feasible[10]. Additionally, many such technologies are also available via various cloud-computing platforms such as Amazon Web Services (AWS), Google Cloud Platform, Cloudera, as well as vendors, such as DNAnexus, BaseSpace, SevenBridges, Cypher Genomics[11,12]. However, these options are often costly and out of reach for the majority of modest size research groups. While cloud computing aids in analytical performance by improving computing time and storage, there is considerable room for improvement in current software design in biomedical research for cloud-based big-data analysis.

The manuscripts in this session highlight the importance of network-based methods in identifying patterns and address the diverse range of challenges associated with machine

learning techniques. The applications of these methods are well demonstrated in EHR, next-generation sequencing data, as well as in simulated datasets as described below.

2. Session Contributions:

2.1 Network-based approaches

Network-based methods for pattern and data mining have gained popularity as efficient computational approaches[13,14]. For example, networks can be used for explaining associations among genetic variants and diseases where diseases and variants are represented as nodes, and associations are represented as edges. Applying various network analysis techniques has also helped in identifying hidden patterns in datasets which are otherwise not visible when results are evaluated in a tabular form[15,16]. The utility of networks is critical in integrating results from various association analyses as well as integrating multi-omic data sets in identifying combinatorial effects of variations on phenotypes. Along with representing associations, networks can also be used in identifying a non-constituent effect of different variables on a phenotype.

In the manuscript titled *"Functional network community detection can disaggregate and filter multiple underlying pathways in enrichment analysis"*, **Harrington** et al. address challenges with identifying pathways from differential expression analyses in non-network based methods. They demonstrate how applying a network based approach that combines community detection with functional networks can help in identifying true positive pathways. They applied the proposed method on simulated dataset and showed its utility on a biological dataset to discover pathways enriched across high grade serous ovarian cancer (HGSC).

Agarwal et al. address the challenge of dealing with imperfect and noisy molecular network data to uncover disease pathways and proteins in their manuscript titled *"Large-Scale Analysis of Disease Pathways in the Human Interactome"*. The authors conducted a comprehensive network analysis on publicly available data from human protein-protein interaction (PPI) network and DisGeNET database containing protein-disease associations. They observed that several proteins associated with a disease tend to fall in different pathways that are not necessarily well connected. This analysis could be useful in the future development of network-based methods to identify robust pathways.

Biomedical data is highly heterogeneous and incomplete, making extraction of meaningful biologically information a major challenge[17,18]. In their manuscript titled *"OWL-NETS: Transforming OWL Representations for Improved Network Inference"*, **Callahan** et al. propose a novel method for abstracting complex, heterogeneous biological knowledge into lossless network representations that facilitate network inference. The OWL-NETS method could help in enhancing network inference where multi-omic and complex biological information is utilized.

2.2 Machine learning approaches

Deep learning and machine learning techniques are an integral component of evaluating biomedical data, and their use has been increasing dramatically over the past decade[19,20]. Machine learning methods are used extensively for identification of correlations between variables, e.g. between different phenotypes, or between phenotypes and genotypes. Moreover, many methods have demonstrated their applications in other-omics datasets such as proteomics, transcriptomics, and metabolomics[21–23]. Methods such as unsupervised learning are independent of any set rules for identifying patterns to associate or correlate variables. These methods have also gained popularity in the field of biomedical data to improve prediction of health outcomes by mining biologically relevant data.

Recently, machine learning methods, both supervised and unsupervised, have been widely used in the field of biomedical informatics. Though the concept of machine learning is not new, researchers still struggle to identify the best method suited for identifying viable solutions to their problems. In their manuscript titled *"Data-driven Advice for Applying Machine Learning to Bioinformatics Problems",* **Olson** et al., present a wide range of comparisons among various machine learning methods, and show how effectively tuning the methods could enable identification of true positive results.

Machine learning methods are also extensively used in analyses of medical imaging data, such us in cancer radiomics, an emerging field focused on quantification of tumor phenotypes using various imaging features. In the manuscript titled *"Tree-based Methods for Characterizing Tumor Density Heterogeneity",* **Shoemaker** et al. propose a novel decision tree-based approach to quantify heterogeneous tumor characteristics from imagining data, using CT scans of solid adrenal lesions as an example.

In the manuscript titled "*Improving the Explainability of Random Forest Classifier – User Centered Approach"*, **Petkovic** et al. propose a novel approach to explain complex models generated from one of the most popular machine learning classifier methods, random forests. Through their method and its application, authors provide an effortless way to generate summary reports of data to enhance the interpretability of complex random forest classifiers.

2.2 Application of methods to identify patterns in EHR data

EHR data consists of a wealth of information about patients. These datasets are present in forms of patient records on disease diagnosis, lab tests which include blood tests as well as imaging data, demographic information, medication information, as well as physicians' clinical notes. Patients' EHRs can be linked with their genetic data in a form of biobanks (for example Geisinger's Mycode Community Health Initiative, Vanderbilt's BioVU, eMERGE Network, UK BioBank)[24–26], which provides a great opportunity for uncovering novel disease associations and, ultimately, improving health care.

EHR data are a great source of phenotype information. The criteria used for assigning a disease status (case and control status) to a patient sample vary greatly across different studies. Some studies use extensive manual curation and development of a phenotypic algorithm to assign patients disease status, whereas other studies use instances of disease diagnosis codes (ICD-9 codes) to assign case-control status to patients[27,28]. High-throughput techniques to generate phenotypes are necessary to bridge the gap between the two techniques described above. In the manuscript titled *"Automated Disease-Cohort Selection using Word Embeddings from Electronic Health Records",* **Glicksberg** et al. address this problem by evaluating the performance of automated feature learning method, word2vec, with the established research-based electronic phenotyping algorithms in extracting cohorts for five diseases.

In manuscript titled *"Mapping Patient Trajectories using Longitudinal Extraction and Deep Learning in the MIMIC-III Critical Care Database"*, **Beaulieu-Jones** et al. apply deep learning techniques to map patient time series data to the preventive care that a patient receives in the EHR. The challenge of utilizing dense longitudinal information from EHR data is addressed in this manuscript. Machine learning methods and the comparisons of methods highlighted in this paper also provide useful insights towards deep learning techniques and their applications in pattern identification.

In the manuscript titled "*Causal Inference on Electronic Health Records to Assess Blood Pressure Treatment Targets: An Application of the Parametric g Formula",* **Johnson** et al. use EHRs to extract longitudinal blood pressure information from patients suffering from hypertension. They use this data to demonstrate the utility of an established causal inference technique, parametric g-formula, for the first time in EHR data in the context of cardiovascular preventative medicine.

2.3 Applications in transcriptome and next-generation sequencing data

With the availability of next-generation sequencing data (NGS), research focused on pattern recognition for identification of functional elements in the human genome has become widespread. Analyzing gene expression information is helpful in understanding the influence of such elements on a trait or disease.

Jeong et al. hypothesized that analyzing transposable elements (TE), which for a long time had been incorrectly labeled as junk DNA, could provide useful functional insights for biomedical data. In their manuscript, *"An Ultra-Fast and Scalable Quantification Pipeline for Transposable Elements from Next Generation Sequencing Data"*, the authors propose a pipeline to quantify TE in the genome from NGS data. This pipeline could be useful for the biomedical informatics community to discover hidden association among TE expression and diseases.

One of the major goals of association analysis is to identify the proportion of phenotypic variance explained by genetic variations. Transcriptome-wide association analysis (TWAS)

are becoming popular methods in explaining the proportion of phenotypic variance that cannot be explained by single nucleotide variations alone[9,29,30]. In the paper titled *"How powerful are summary-based methods for identifying expression-trait associations under different genetic architectures?",* **Veturi** et al. use a simulation study to analyze two major approaches for conducting TWAS, TWAS-MP (multi-SNP prediction) and TWAS-SMR (summary-based Mendelian Randomization). The paper describes a comprehensive power analysis for detecting gene-trait analysis, which in the future could be expanded to other kinds of omics datasets.

3. References

1. Pattern Recognition and Machine Learning | Christopher Bishop | Springer.
2. Minelli, M., Chambers, M. & Dhiraj, A. Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses. (John Wiley & Sons, 2012).
3. Iddamalgoda, L. et al. Data Mining and Pattern Recognition Models for Identifying Inherited Diseases: Challenges and Implications. Front. Genet. **7,** (2016).
4. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. **19,** A68–A77 (2015).
5. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature **526,** 68–74 (2015).
6. Sudlow, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. **12,** e1001779 (2015).
7. Mailman, M. D. et al. The NCBI dbGaP database of genotypes and phenotypes. Nat. Genet. **39,** 1181–1186 (2007).
8. Bourne, P. E. et al. The NIH Big Data to Knowledge (BD2K) initiative. J. Am. Med. Inform. Assoc. JAMIA **22,** 1114 (2015).
9. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. **18,** 117–127 (2017).
10. Hadoop & Spark | IBM Big Data & Analytics Hub. Available at: /technology/hadoop-and-spark. (Accessed: 20th September 2017)
11. Hashem, I. A. T. et al. The rise of "big data" on cloud computing: Review and open research issues. Inf. Syst. **47,** 98–115 (2015).
12. Big Data is a Big Deal for Biomedical Research. whitehouse.gov (2013). Available at: https://obamawhitehouse.archives.gov/blog/2013/04/23/big-data-big-deal-biomedical-research. (Accessed: 20th September 2017)
13. Silva, T. C. & Zhao, L. Network-based high level data classification. IEEE Trans. Neural Netw. Learn. Syst. **23,** 954–970 (2012).
14. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. Bioinforma. Oxf. Engl. **27,** 431–432 (2011).
15. Barabási, A.-L. Network Medicine — From Obesity to the "Diseasome". N. Engl. J. Med. **357,** 404–407 (2007).
16. Goh, K.-I. et al. The human disease network. Proc. Natl. Acad. Sci. **104,** 8685–8690 (2007).
17. Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P. & Burgun, A. Integrating Heterogeneous Biomedical Data for Cancer Research: the CARPEM infrastructure. Appl. Clin. Inform. **7,** 260–274 (2016).

18. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the Missing Link for Big Biomedical Data. JAMA **311,** 2479–2480 (2014).
19. Deng, L. & Yu, D. Deep Learning: Methods and Applications. Found. Trends® Signal Process. **7,** 197–387 (2014).
20. Special Issue: Deep Learning for Biomedical and Health Informatics. Available at: https://jbhi.embs.org/2016/12/30/special-issue-deep-learning-biomedical-health-informatics/. (Accessed: 20th September 2017)
21. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene-gene interactions: a review. Appl. Bioinformatics **5,** 77–88 (2006).
22. Zheng, T. et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int. J. Med. Inf. **97,** 120–127 (2017).
23. 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. **16,** 85–97 (2015).
24. Gottesman, O. et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. Off. J. Am. Coll. Med. Genet. (2013). doi:10.1038/gim.2013.72
25. Carey, D. J. et al. The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. (2016). doi:10.1038/gim.2015.187
26. Development of a large-scale de-identified DNA biobank to enable personalized medicine. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/18500243. (Accessed: 20th September 2017)
27. Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. Nat. Rev. Genet. **17,** 129–145 (2016).
28. Kirby, J. C. et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. J. Am. Med. Inform. Assoc. JAMIA **23,** 1046–1052 (2016).
29. Gusev, A. et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. Am. J. Hum. Genet. **95,** 535–552 (2014).
30. Lonsdale, J. et al. The Genotype-Tissue Expression (GTEx) project. Nat. Genet. **45,** 580–585 (2013).