Pacific Symposium on Biocomputing 25:286-294(2020)

Ongoing challenges and innovative approaches for recognizing patterns across large-scale, integrative biomedical datasets

Shilpa Nadimpalli Kobren
Dept. of Biomedical Informatics, Harvard Medical School
10 Shattuck Street, 4th Floor
Boston, MA 02155
Email: shilpa_kobren@hms.harvard.edu

Brett Beaulieu-Jones
Dept. of Biomedical Informatics, Harvard Medical School
10 Shattuck Street, 4th Floor
Boston, MA 02155
Email: brett_beaulieu-jones@hms.harvard.edu

Christian Darabos
Research, Teaching, and Learning at IT&C, Dartmouth College
37 Dewey Field Road
Hanover, NH 03755
Email: christian.darabos@dartmouth.edu

Dokyoon Kim
Dept. of Biostatistics and Epidemiology, Perelman School of Medicine at UPenn
3400 Civic Center Boulevard, Building 421
Philadelphia, PA 19104
Email: dokyoon.kim@pennmedicine.upenn.edu

Anurag Verma
Dept. of Genetics, Perelman School of Medicine at UPenn
3400 Civic Center Boulevard, Building 421
Philadelphia, PA 19104
Email: anuragv@upenn.edu

1 Introduction

The size, complexity, and heterogeneity of biologically and medically relevant datasets have been dramatically increasing in recent decades in conjunction with newly developed experimental and clinical technologies. Innovative computational approaches have been essential in processing these datasets in order to glean the patterns and insights necessary to support translational research (Bourne et al., 2015). This combination of increasing data availability with novel method development has the potential to

greatly advance the field of biomedicine, but there are many unique challenges to overcome first. When developing and employing computational tools to extract meaningful patterns from biomedical data, sophisticated algorithms are required to handle the pervasive inconsistencies, sparseness, and noisiness unique to biomedical datasets. Clinical data, contained most commonly in electronic patient medical health records, can be analyzed in aggregate to uncover patterns that highlight, for instance, complex disease traits, disease onset, or drug reactions that would otherwise have gone unnoticed. Moreover, advancements in high-throughput molecular assays have generated swathes of multi-omic biological data, including genomic, proteomic, transcriptomic and epigenomic data. The decrease in cost of sequencing and related molecular assays has resulted in these tests becoming more routine in the clinic and thus has led to rapidly expanding sets of molecular omics data available for analysis. The integration of clinical data with these omics data and subsequent analysis of heterogeneous datasets is nontrivial and often requires modern statistical, machine learning approaches (Ritchie et al., 2015). The fitting of such models and analysis of their results promises the opportunity for an increased understanding of complex phenotypic traits associated with healthy individuals, those with disease, and the spectrum in between (Bersanelli et al., 2016). Here, we will highlight and describe recent, innovative ways to address the challenges associated with deriving meaningful patterns from biomedical data, including dealing with the quantity, quality, and integration of data.

Cutting edge research discussed here can be broadly categorized into four main areas: (i) identifying patterns in clinical data (e.g., electronic health records or pathology images), (ii) utilizing and integrating molecular omics data (e.g., genome sequencing and gene expression data), (iii) statistical machine learning and graph-based method development, and (iv) computational challenges and reproducibility.

2 Identifying patterns in clinical data

One of the most comprehensive forms of clinical data are electronic health records (EHRs), which contain longitudinal health information for an individual patient at the point of care. EHRs include extensive yet noisy clinical notes, diagnosis and procedure codes, medication prescriptions, laboratory test measurements and results as well as pathology images where relevant. There has been a litany of observational or retrospective case-control studies that demonstrate the potential for using EHR data to improve our understanding of disease risk and prevalence across the general population. Moreover, automated analyses of histopathology images in conjunction with other clinical information can improve disease diagnosis and stratification. However, drawing conclusions from joint analyses of EHR data requires careful consideration of the biases inherent in EHR, including missing data, institutional discrepancies in coding practices, and high-throughput electronic phenotyping.

In their paper entitled “Clinical concept embeddings learned from massive sources of multimodal medical data,” Beam et al. take advantage of extremely large medical data to derive clinical concept embeddings. These concept embeddings are a mathematical representation (i.e., a vector) of clinical events, where similar clinical events will have similar representations. These vectors are learned from patient medical records and can be subsequently used to make comparisons between patients and to use as input for machine learning models. Similarly, in the manuscript “Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular

data," Neums et al. perform data integration across different data modalities (e.g., clinical records and molecular datasets including gene expression) to improve prognosis predictions for kidney, lung and bladder cancers.

One particularly rich source of information contained in medical records are pathology images. Pathology refers to the microscopic examination of body tissue for the purpose of diagnosis, to study the manifestation of known diseases, or for forensic analysis. Images of body tissue used by pathologists are included in EHRs, and though they provide highly disease-relevant information, they often require processing by domain experts to be used for translational research. Methods that automatically derive patterns from sets of pathology images are critical for reducing human error and improving diagnosis and disease stratification. In this session, Levy et al. will describe PathFlowAI, their automated and efficient workflow for processing and performing analytics and interpretation of digital pathology images. In the same vein, Hao et al. will present PAGE-Net, a deep learning approach that integrates histopathological images with genomic data for an interpretable analysis of survival.

Beyond analyzing images at a microscope resolution, electron microscopy techniques enable researchers to observe molecular events at an intensely greater resolution. Gur et al. analyze temporal (4D) electron microscopy data to uncover microvascular dynamics using a temporal segmentation approach. As illustrated by these authors, the utility of temporal electron microscopy data can only be fully realized with advanced computational techniques for pattern elicitation.

3 Integration and analysis of molecular omics data

The dramatic drop in the cost of clinical sequencing has resulted in an unprecedented number of sequenced genomes available for analysis. These data include complete genome sequences for healthy patients as well as complete genome sequences for those with disease, "somatic" mutations in cancer patients derived from a genetic comparison between matched normal and tumor tissue samples, and transcriptome sequences for specific tissue types and even single cells. In addition, other forms of omics "big data" stem from high-throughput assays measuring protein-protein interactions, protein-DNA interactions, gene expression, and downstream functional molecular effects. Finding patterns across these large-scale omics data in order to increase our understanding of the molecular processes underlying human diseases remains a challenging task (Pasaniuc and Price, 2017).

Novel work presented in this session pertaining to the integration and analysis of molecular omics data falls primarily into two areas: (i) data analysis to further our understanding of transcriptional and post-transcriptional processes and (ii) uncovering genes or pathways relevant to cancer onset and progression.

3.1 Transcriptomics

Transcription refers to the process where subsets of genes encoded in our genomes are "turned on" to produce corresponding RNA strands; these RNA can then code for protein molecules or directly carry out a variety of functions in the cell. Differing sets of genes are transcribed across different cell types, at different points in development, or in response to different environmental stimuli. Quite often, the

dysregulation of this process is linked to human diseases. As such, experimental assays to determine which genes are being transcribed across different tissues or in different individuals have been essential in uncovering key characteristics of disease.

One of the issues with these experimental assays, for instance microarrays or RNA sequencing, is that the level of RNA transcripts can be quite low, and therefore data for certain lowly-transcribed genes can often be missing from experimental results. This biases all downstream analyses and can result in critical associations being missed. Various imputation methods now exist for filling in this missing gene expression data before further analysis. Bobak et al. present a thorough assessment of several state-of-the-art imputation methods for gene expression through a meta-analysis of distinct cohorts of tuberculosis patients. In addition to data sparsity, another issue with transcriptional data is its heterogeneity. In their paper, Tran et al. develop a way to represent the functions of heterogeneous RNA sequences by integrating this data with various multi-omics, interaction and annotation data.

Even after genes are transcribed, there are various additional cellular processes involving RNA transcripts, such as RNA degradation, that can further regulate gene expression. These processes, collectively referred to as "post transcriptional regulation," remain poorly understood, due in large part to difficulties in developing experimental techniques to directly measure relevant events. To alleviate this issue, Srivastava et al. present PTR Explorer, a new approach that uses proteogenomics data to identify and further explore post transcriptional regulatory mechanisms.

3.2 Cancer driver detection

As mentioned, dysregulation of gene transcription is often linked to human disease. One disease characterized by extensive gene dysregulation is cancer. Despite numerous efforts over the past decade, comprehensively identifying the set of genes and pathways with roles in cancer onset and progression (i.e., "drivers") and the mode of action for specific mutations within these genes and pathways has yet to be achieved. Haan et al. present a novel way to directly utilize transcriptional signatures in order to find cancer drivers with their method LURE. Park et al. take a different approach toward a similar goal of finding pathways associated with cancer phenotypes; their method uses Bayesian semi-nonnegative matrix tri-factorization.

Tumors have several signatures which can be detected through various experimental means. As previously mentioned, the expression of genes within tumors is distinct from the expression of genes in matched normal tissue. In addition, tumors have several somatic mutations, some of which occur in frequently mutated genes across patient cohorts. Histopathological images of cancer tissues have several features that are not present in normal human tissue. More recently, researchers have found that epigenetic traits, including DNA hypermethylation and accessibility, are also different in tumor genomes. Comparing these different sources of data–all of which provide evidence that can be used for cancer detection and cancer type stratification–is computationally and conceptually difficult. In their paper, Kompa and Coker present a step toward data integration by learning a latent space of highly multidimensional cancer data.

4 Interpretable and graph-based machine learning approaches

Statistical machine learning approaches have long been applied in the biomedical problem-space. Of the research presented here with a focus on method evaluation and development, there are two distinct trends: (i) interpretability of machine learning models to evaluate patterns in computational biology, and (ii) graph-based approaches to prioritize and understand relationships between concepts.

4.1 Machine learning to support the interpretability of complex relationships

Statistical, machine learning approaches are commonly applied for pattern recognition across biomedical datasets. The input to these methods is the data (across which patterns should be derived) in the form of an ordered list of comparable values (e.g., numbers). There are several approaches for predicting classes or outcomes by statistically detecting patterns across the input data. However, although these methods often perform quite well, they also inherently obscure the contribution of features in the input data, making it difficult to interpret resultant models in order to derive clinically or medically relevant insights. Developing more accurate models in a domain-specific context as well as increasing the interpretability of these models are two important goals that are addressed here.

One type of machine learning classifier is a "random forest," where feature values that can discern between classes in the input data are represented as nodes in a collection of decision trees. In their paper entitled "Tree-weighting for multi-study ensemble learners," Ramchandran et al. consider novel ways to weight different trees before combining their predictions in an ensemble approach. Importantly, they explore how these weights correspond to the tree structure, enabling interpretation of which features contribute to the weighting process.

There are also domain-specific applications of machine learning approaches that require novel method development. For instance, Stanley et al. aim to predict which sites within human proteins are susceptible to being cleaved by the dengue viral protease. In their paper, "Two-stage ML classifier for identifying host protein targets of the dengue protease," the authors describe how to extract biochemical and protein secondary structure features from their input data to eventually build a highly accurate predictor; the eventual usage of such features by their classifier can potentially be analyzed to learn more about the dengue protease. Hocking and Bourque address a different biological problem, that of jointly detecting "peaks" in epigenomic data analysis. More specifically, the binding of transcription factor proteins to the genome to affect gene expression is a form of epigenetic regulation; determining the location of these binding sites requires the detection of "peaks" of overlapping mapped DNA sequencing reads. The authors' method PeakSegPipeline can be used to detect multiple peaks jointly, and importantly, uses a more interpretable model to uncover peaks that overlap at the same positions.

Lee et al. address yet another biologically relevant problem, the inference of gene regulatory networks using gene expression data. The sparseness and noisiness of gene expression data was discussed in Section 3.1, and the utility of biologically networks, including transcriptome networks, is further discussed in the next section. In their paper, the authors present a new algorithm, NO-BEARS, that dramatically reduces the computational complexity of previous approaches for this task and, by utilizing GPU computation, drops the required compute time from hours to seconds.

4.2 Graph-based approaches to prioritize and understand relationships

Graphs or networks, defined by a set of nodes connected by edges, are a powerful mathematical concept that have been repeatedly utilized in the fields of computational biology and biomedical informatics to model complex relationships between entities and decipher patterns across these entities (Cowen et al., 2017). In this session, a number of manuscripts use graphs for various applications in biomedical informatics.

First, molecular omics data can often be represented as a graph, most commonly as protein–protein interaction networks. Pham and Lichtarge present a new method for an old problem in their manuscript entitled "Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks." Yao and Ramsey derive features from the same types of molecular networks used by Pham and Lichtarge, and use these features to improve the computational prediction of functional noncoding single-nucleotide polymorphisms (SNPs) in their method, CERENKOV3. Indeed, understanding the functional importance of SNPs that fall outside of genes remains an extremely challenging problem in computational biology. Many of these SNPs may have functionality with low penetrance, where their true effect can only be measured in the context of several other genetic changes. Validating the functionality of noncoding SNPs has also proved notoriously difficult.

It is thought that several rare or undiagnosed diseases might have etiological origins from changes within noncoding parts of the genome. As mentioned, identifying such mutations is quite challenging. However, finding innovative ways to treat these disorders before having a complete understanding of their molecular basis would be highly impactful. Sosa et al. present a graph-based method for exactly this purpose, and present their findings in their manuscript, "A literature-based knowledge graph embedding method for identifying drug repurposing opportunities for rare diseases."

5 Computational challenges and reproducibility

Moving forward, one of the biggest challenges across several scientific fields including the field of biomedicine is that of reproducibility. Biases, inconsistencies and general shift over time across datasets can severely impact results and lead to erroneous conclusions or pattern detection. Aggregating and integrating datasets properly can reduce the prevalence of this issue. To enable this, however, computational methods must be scalable and allow for real-time dataset changes and additions. Various cutting-edge solutions for this problem are presented in this session. First, Wheeler et al. discuss and evaluate solutions for reproducing and scaling large-scale studies that analyze genomic sequences. Similarly in their manuscript, Wang et al. address the issue of reproducibility through the lens of interpreting and assessing accuracy of prediction methods. Their work enhances the interpretability and accuracy of a learned model for predicting disease progression by using phenotype-based patient similarity. Finally, incorporating and integrating novel data in a machine learning model can be critical for confirming trends that were previously established using alternate datasets. In their paper, Burkhardt et al. address the serious issue of discovering side effects of drugs by using active learning and crowdsourcing to analyze social media data.

6 Discussion

The exciting new work we have briefly summarized here touches on a breadth of topics related to the computational processing of "big" clinical, biomedical, and omics datasets with the goal of recognizing patterns relevant for translational research.

One theme that recurs throughout these papers is the multimodality and heterogeneous nature of biomedical data. Whereas some papers address this issue as their primary focus (e.g., Beam et al., Tran et al.), other papers deal with this issue indirectly by successfully processing and integrating diverse datasets with a distinct end goal. The heterogeneity of data stems in part from differing data types, but also arises through institutional or technical artifacts. Indeed, as novel technologies enable more clinics, hospitals, and research centers to integrate and access clinical records across sites, for instance through Fast Healthcare Interoperability Resources (FHIR) or the Observational Medical Outcomes Partnership (OMOP), heterogeneity stemming from institutional artifacts will continue to grow but may become more transparent.

Another theme that threads through the papers in this session is that of innovation in the algorithmic and methods space. Several machine learning methods can be applied on new datasets and toward new applications. However, sometimes attempts to apply existing methods to new problem areas generates enough complications to warrant the development of entirely new methods. Several new algorithmic methods were developed to address pertinent issues in the biomedical field, including interpreting transcriptomes (e.g., Lee et al., Srivastava et al.), processing image data (e.g., Levy et al., Hao et al.), and deriving features from molecular networks (e.g., Pham and Lichtarge, Yao and Ramsey). The innovative development of new methods is essential for furthering the field of biomedical informatics and enabling researchers now and in the future to continue to recognize important patterns throughout biomedical data.

References

AL Beam, B Kompa, A Schmaltz, I Fried, G Weber, N Palmer, X Shi, T Cai, IS Kohane (2019). "Clinical concept embeddings learned from massive sources of multimodal medical data." *Pac Symp Biocomput.* To appear.

M Bersanelli, E Mosca, D Remondini, E Giampieri, C Sala, G Castellani, L Milanesi (2016). "Methods for the integration of multi-omics data: mathematical aspects." *BMC Bioinformatics,* 17:S15. doi:10.1186/s12859-015-0857-9

CA Bobak, L McDonnell, MD Nemesure, J Lin, JE Hill (2019). "Assessment of imputation methods for missing gene expression data in meta-analysis of distinct cohorts of tuberculosis patients." *Pac Symp Biocomput.* To appear.

PE Bourne, V Bonazzi, M Dunn, ED Green, M Guyer, G Komatsoulis, J Larkin, B Russell (2015). "The NIH Big Data to Knowledge (BD2K) initiative." *J Am Med Inform Assoc,* 22:1114.

S Burkhardt, J Siekiera, J Glodde, MA Andrade-Navarro, S Kramer (2019). "Identifying drug side effects from social media using active learning and crowd sourcing." *Pac Symp Biocomput.* To appear.

L Cowen, T Ideker, BJ Raphael, R Sharan (2017). "Network propagation: a universal amplifier of genetic associations." *Nat Rev Genet,* 18(9):551-562. doi:10.1038/nrg.2017.38

S Gur, L Wolf, L Golgher, P Blinder (2019). "Microvascular dynamics from 4D microscopy using temporal segmentation." *Pac Symp Biocomput.* To appear.

D Haan, R Tao, V Friedl, IN Anastopoulos, CK Wong, AS Weinstein, JM Stuart (2019). "Using transcriptional signatures to find cancer drivers with LURE." *Pac Symp Biocomput.* To appear.

J Hao, SC Kosaraju, NZ Tsaku, DH Song (2019). "PAGE-Net: Interpretable and integrative deep learning for survival analysis using histopathological images and genomic data." *Pac Symp Biocomput.* To appear.

TD Hocking, G Bourque (2019). "Machine learning algorithms for simultaneous supervised detection of peaks in multiple samples and cell types." *Pac Symp Biocomput.* To appear.

B Kompa, B Coker (2019). "Learning a latent space of highly multidimensional cancer data." *Pac Symp Biocomput.* To appear.

H-C Lee, M Danieletto, R Miotto, ST Cherng, JT Dudley (2019). "Scaling structural learning with NO-BEARS to infer causal transcriptome networks." *Pac Symp Biocomput.* To appear.

JJ Levy, LA Salas, BC Christensen, A Sriharan, LJ Vaickus (2019). "PathFlowAI: A convenient high-throughput workflow for preprocessing, deep learning analytics and interpretation in digital pathology." *Pac Symp Biocomput.* To appear.

L Neums, R Meier, DC Koestler, JA Thompson (2019). "Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data." *Pac Symp Biocomput.* To appear.

S Park, N Kar, TH Hwang (2019). "Bayesian semi-nonnegative matrix tri-factorization to identify pathways associated with cancer phenotypes." *Pac Symp Biocomput.* To appear.

B Pasaniuc, AL Price (2017). "Dissecting the genetics of complex traits using summary association statistics." *Nat Rev Genet,* 18:117-127. doi:10.1038/nrg.2016.142

M Pham, O Lichtarge (2019). "Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks." *Pac Symp Biocomput.* To appear.

M Ramchandran, P Patil, G Parmigiani (2019). "Tree-weighting for multi-study ensemble learners." *Pac Symp Biocomput.* To appear.

MD Ritchie, ER Holzinger, R Li, SA Pendergrass, D Kim (2015). "Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet,* 16:85-97. doi:10.1038/nrg3868

DN Sosa, A Derry, M Guo, C Brinton, E Wei, RB Altman (2019). "A literature-based knowledge graph embedding method for identifying drug repurposing opportunities for rare diseases." *Pac Symp Biocomput.* To appear.

A Srivastava, M Sharpnack, K Huang, P Mallick, R Machiraju (2019). "PTR Explorer: An approach to identify and explore Post Transcriptional Regulatory mechanisms using proteogenomics." *Pac Symp Biocomput.* To appear.

JT Stanley, AR Gilchrist, AC Stabell, MA Allen, SL Sawyer, RD Dowell (2019). "Two-stage ML classifier for identifying host protein targets of the dengue protease." *Pac Symp Biocomput.* To appear.

N Tran, J Gao (2019). "Functional representation of large-scale heterogeneous RNA sequences with integration of diverse multi-omics, interactions, and annotations data." *Pac Symp Biocomput.* To appear.

Y Wang, T Wu, Y Wang (2019). "Enhancing model interpretability and accuracy for disease progression prediction via phenotype-based patient similarity learning." *Pac Symp Biocomput.* To appear.

NR Wheeler, P Benchek, BW Kunkle, KL Hamilton-Nelson, M Warfe, JR Fondran, JL Haines, WS Bush (2019). “Big data solutions for reproducibility and scalability of genomic sequencing studies.” *Pac Symp Biocomput.* To appear.

Y Yao, S Ramsey (2019). “CERENKOV3: Clustering and molecular network-derived features improve computational prediction of functional noncoding SNPs.” *Pac Symp Biocomput.* To appear.