DISCOVERY OF MOLECULARLY TARGETED THERAPIES

KELLY REGAN
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Kelly.Regan@osumc.edu

ZACHARY ABRAMS
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Zachary.Abrams@osumc.edu

MICHAEL SHARPNACK
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Michael.Sharpnack@osumc.edu

ARUNIMA SRIVASTAVA
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
srivatava.1@osu.edu

KUN HUANG
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Kun.Huang@osumc.edu

NIGAM SHAH
Center for Biomedical Informatics Research
Stanford University
Stanford, CA 94305
nigam@stanford.edu

PHILIP R.O. PAYNE
Department of Biomedical Informatics
The Ohio State University
Columbus, OH 43210
Philip.Payne@osumc.edu

1. Introduction

The delivery of personalized healthcare is predicated on the application of the best available scientific knowledge to the practice of medicine in order to promote health, improve outcomes and enhance patient safety [1-3]. Unfortunately, current approaches to basic science research and clinical care are poorly integrated, yielding clinical decision-making processes that do not take advantage of up-to-date scientific knowledge [2-4]. Basic scientists investigating the biological basis for a given disease may regularly encounter synergistic effects spanning two or more bio-molecular entities or processes that can contribute to our understanding of the mechanisms underlying phenomena such as the etiologic basis of the targeted disease state or potential response to therapeutic agents [5]. However, systematic approaches to the use of that knowledge in order to directly inform the selection of targeted molecular therapies for "real world" patients are extremely limited [1, 3, 6-9]. There are an increasing number of multi-modelling and in-silico knowledge synthesis techniques that can provide investigators with the tools to quickly generate hypotheses concerning the relationships between entities found in heterogeneous collections of scientific data — for example, exploring potential linkages among genes, phenotypes and molecularly targeted therapeutic agents, thus enabling the "forward engineering" of treatment strategies based on knowledge generated via basic science studies [1, 4, 6, 10, 11]. Ultimately, the goal of such methodologies is to accelerate the identification of actionable research questions that can make direct contributions to clinical practice. Given increasing concerns over the barriers to the timely translation of discoveries from the laboratory to the clinic or broader population settings, such high-throughput hypothesis generation and testing is highly desirable [1, 4, 6, 8, 12]. These needs are particularly critical in numerous disease areas where the availability of new therapeutic agents is constrained, thus calling for the re-use and repositioning of existing treatments [13, 14].

In response to the challenges and opportunities enumerated above, there exits an emerging body of research and development focusing on multi-modeling approaches to the discovery of molecularly targeted therapies, including experimental paradigms spanning a spectrum from the identification of molecular targets for drugs, to the repurposing or repositioning of existing agents that utilize such targets, to the systematic identification of novel combination therapy regimens that amplify or enhance the effectiveness of their constituent components. This focus is motivated by recent and significant advances in the state of systems biology and medicine that have demonstrated that the ability to generate and reason across complex and scalar models is essential to the discovery of high-impact biologically and clinically actionable knowledge [1, 4, 12]. Such approaches are designed to overcome the limitations of reductionist approaches to scientific discovery, replacing decomposition-focused problem-solving with integrative network-based modeling and analysis techniques [4, 8]. Systems-level analysis of complex problem domains ultimately enables the study of critical interactions that influence health and wellness across a scale from molecules to populations, and are not observable when such systems are broken down into constituent components. The use of systems-level analysis methodologies is well supported by the foundational theory of vertical reasoning first proposed by Blois [15]. This theory holds that effective decision-

making in the biomedical domain is predicated on the vertical integration of multiple, scalar levels of reasoning. This fundamental premise is the basis for a correlative framework put forth by Tsafnat and colleagues, which states that the ability to replicate expert reasoning relative to complex biomedical problems using computational agents (e.g., in-silico knowledge synthesis) requires the replication of such multi-scalar and integrative decision-making [16]. In order to achieve such an outcome, Tsafnat posits that multi-scalar decision-making in an in-silico context requires both: 1) the generation of component decision-making models at multiple scales; and 2) the similar generation of interchange layers that define important pair-wise connections between entities situated in two or more component models, often referred to as vertical linkages [16]. When such component models and interchange layers are combined in a computationally actionable format, they yield what can be referred to as a multi-model for a given domain that is able to satisfy the premises of Blois' vertical reasoning axiom, and therefore facilitate the replication of expert performance in a high-throughput manner [16]. Of note, this type of approach is extremely reliant upon graph-theoretic reasoning and representational models, using a network paradigm that allows for the application of logical reasoning operations spanning the entities and relationships that make up a multi-model [8]. Network paradigms have been regularly shown to be the ideal representational model for naturally occurring systems, such as the 'scale-free' networks encountered in biological and clinical phenomena [8]. At the most basic level, network-based multi-modeling across scales presents an elegant and computationally tractable approach to understanding and evaluating complex biological and clinical systems in order to discover the knowledge incumbent to such constructs. This type of approach benefits from a robust set of foundational theories and frameworks that can inform and shape the application of multi-modeling techniques to a variety of knowledge discovery use cases. As such, there is a growing body of evidence concerning the application of network-based approaches to multi-modeling with an emphasis on therapeutic agent discovery, re-positioning and molecular targeting. Examples of such evidence include reports and perspectives published by Hood and Perlmutter [1], Butcher and colleagues [12], and Lussier and Chen [13].

2. Overview of Session Contributions

The utility and impact of multi-modeling approaches to integrative biological and clinical analyses, including hypothesis discovery operations such as those related to the identification of molecularly targeted therapies as noted above, have been explored in a number of instances by the biology, computer science and translational bioinformatics communities. At a high level, the exemplary efforts made by authors contributing to this session of PSB 2016 provide a broad cross-section of such novel methods, and focus on: 1) the development of factorization-based models to traverse multiple large-scale database comprising types of drug-disease and drug-target relationships (**Zitnik *et al*** and **Regenbogen *et al***); 2) network-theoretic approaches in a variety of applications including: linking environmental risk factors for disease via systematic analysis of biological pathways (**Darabos *et al***), the prioritization of gene mutations causing drug resistance (**Verkhivker**), and the facilitation of viable community detection (**Yu *et al***); and 3) the incorporation of prior knowledge into in silico methods in order to optimize

large-scale regression-based association studies (**Verma *et al***) and to discover dependencies between genes differ across disease conditions (**Speyer *et al***). Brief synopses of these reports are provided below:

2.1 Factorization-based Models for Traversing Databases

Zitnik *et al* describe a novel collective pairwise classification (COPACAR) model for analysis of multi-relational data, including clinical manifestations of diseases, molecular interactions of diseases, drug-drug and drug-target interactions and drug-drug similarities. Their model combines factorization models that are optimized for large relational data with classification pairwise ranking loss for classification. Importantly, their model incorporates prior knowledge that is also scalable to highly complex, large-scale data. The authors address the issue of ranking in their predictions, where relationship are ranked according to their relevance, which is ideal for prioritizing large-scale, diverse relationships. They distinguish their approach to other widely recognized collective relational learning approaches optimized to minimize error rate are not well-suited to rank high-confidence relationships integral to applications of precision medicine and drug repurposing. The COPACAR method optimizes a ranking metric using pairwise classification in order to estimate latent factors of entities, which are use to parameterize the model's predictions about pairwise entity relationships. Another particularly significant contribution is the implementation of an application the authors term "category-jumping," which permits the generation of novel hypotheses relating heterogeneous biomedical entities that may be unrecognized by other models that rely on data of a single relation type. The authors demonstrated a widely observed phenomenon that shared clinical manifestations of disease, in particular high-level symptom characteristics, indicate shared molecular interactions (e.g. genetic associations and protein interactions). Finally, hierarchical clustering of the disease matrix demonstrated that diseases with sparse molecular information could be grouped to disease with molecular-rich relations based on clinical manifestations, thus resulting in novel hypotheses for molecular basis of these diseases.

Regenbogen *et al* address an important problem of extrapolating knowledge across diverse, large-scale sources for small-scale, high-resolution problems in personalized medicine, including individual patient drug prediction and drug repositioning. The authors employed a technique called collaborative filtering (CF), which is extensively used in online recommendation systems. Specifically, non-negative matrix factorization (NMF) was used to analyze knowledge of connections, rather than entity features, in order predict interactions among chemicals, genes, and diseases contained within the Comparative Toxicogenomics Database (CTD). Although NMF has been widely used in the analysis of genomics data and for predicting protein-protein and drug-target interactions, a particular novel contribution of this work is the authors' integration across multiple entity types. One benefit of this framework is that it can be easily extended to new entity classes without extensive pre-processing or abstraction, unlike other methods highly specific to entity attributes; however, it is limited to predict interactions among entities without details regarding how entities interact (e.g. directionality, causality, etc.). Their method was able to accurately predict protein-protein interactions in an

independent database and successfully predict CTD entity relationships between successive versions of the database. Furthermore, integrating data across these two independent databases increased the performance of the CF method. Importantly and similar to Zitnik *et al*, the authors confirmed a high degree of precision in their results in addition to a high sensitivity, which is crucial to precision medicine and drug repurposing initiatives that focus on pursuing a small number of hypotheses relative to the total interaction space.

2.2 Network-theoretic Analyses

Darabos *et al* presents a methodology for determining the effect of environmental factors in complex diseases. This is an important problem to address since it is often difficult to distinguish environmental causality in disease development. The authors utilize a tripartite network linking diseases, environmental chemicals and biological pathways in order to identify potential biological effects of environmental chemicals relating to disease. This tripartite network allows for the connecting environmental factors with disease through shared biological processes. The utility of this model is demonstrated in one instance through the linkage of arsenic to multiple diseases through its role in disrupting signal transduction pathways. Overall, this work supports the use of multi-modeling network approaches to elucidate the effects of environmental exposure related to disease states. The authors also show how linking disparate datasets together can help answer large-scale questions through creation of a hypothesis generating system that can help fuel future research areas such as population health and epigenetics.

Verkhivker investigates mechanisms of resistance to lapatinib caused by EGFR mutations. Using genetic and structural data, they are able to prioritize mutations by their ability to affect a residue interaction network, computed using molecular dynamics simulations. The centrality of the residue in the network predicts its ability to disturb the effect of EGFR inhibition. Their results provide a framework for understanding the spectrum of resistance causing mutations, with the added benefit of implying causality of the associated mutations. They suggest that a wide range of mutations within the EGFR protein could cause resistance to lapatinib therapy. Their simulations also recover known resistance mutations, further validating the success of their method.

Yu *et al* propose innovative extensions on the Markov clustering methodology for community detection in networks. While viable community detection has implications in a variety of fields, the authors propose an integrative methodology that is especially apt for garnering a holistic picture in biological networks. They propose two subsequent extensions to the well-known Markov Clustering and regularized Markov Clustering algorithms in order to, firstly, focus on information or influence flow in a non-exclusive manner (inverse regularized Markov Clustering – irMCL) and subsequently integrating network structure with node attributes of biological significance such as phenotype, gene expression or demographic information (attribute inverse regularized Markov Clustering-airMCL). The authors have ideated a method which allows for node attributes to be incorporated in the community detection paradigm, utilizing and weighing attributes with respect to their effect on inter and intra community information flow. They have modeled

the connections between node attributes and network structure in way that is malleable with statistical classification approaches. They prove the validity and robustness of this method by employing it on a simulated as well as real world dataset, utilizing the requisite statistical models and measures for rigor. Their results showcase that the methodology was immune to weak attributes whereas attribute similarity that predicted the structure was highlighted. This eliminates the need for a user-based selection of attribute importance. In the real world Breast Cancer dataset, the algorithm was able to isolate a variety of pathways, including, but not limited to the cell cycle pathway, signal transduction pathway and ribosome biogenesis. Also, the modules isolated showed significant association with time to survival. The authors have aimed to examine and stratify attribute impact by its connection to network structure. This is a novel ideology that promotes multi-modal data integration without succumbing to formation of overly complex models. Finally, with the inclusion and use of classification methodologies in community detection, the authors plan to utilize the inherent classification properties to better select models and features for future work.

2.3 Incorporation of Prior Knowledge into in silico Methods

Verma ***et al*** describe a system of discovering associations utilizing a novel method called Phenome-wide interaction study (PheWIS), which builds on the authors' previous work with phenome-wide association studies (PheWAS). This work seeks to address the problem of discovering associations between single nucleotide polymorphisms (SNPs) and phenotypes on a large scale. The authors approach this problem of large-scale association assessment by modeling the variance of the SNPs. They identified genetic variants that are associated with multiple phenotypes by prioritizing previously published results from both genome-wide and phenome-wide association studies using the AIDS Clinical Trials Group (ACTG) and the Roadmap Epigenome project. They discovered that by filtering out variance from low functional regions of the genome they could conduct a pair-wise search using linear regression analysis to identify associations. With their system the authors were able to identify 50,798 statistically significant associations related to 26 different phenotypes. This work helps to demonstrate not only the importance of modeling genotypic and phenotypic information together but also shows the strength of utilizing previously published information to help inform novel hypothesis driven systems.

Speyer ***et al*** investigate the effect of injecting biological knowledge into a previously developed method, Evaluation of Differential Dependency (EDDY). Their method seeks to answer the question, how do dependencies between genes differ across conditions? They apply their method to the TCGA glioblastoma multiforme data, to find differential dependencies between proneural and nonproneural, and mesenchymal and non-mesenchymal tumors. The result is a list of gene sets whose dependencies most differ between two cancer subgroups. Specifically, they find that the mesenchymal subset is defined by changes to metabolic processes and the proneural subset is defined by changes to AKT-ERK signaling. These pathways are strongly implicated in cancer, which shows the power of this method to find cancer-related results. They compare their results to knowledge-fused differential dependency network (KDDN) and find that the EDDY

method appears to be more sensitive to differential dependencies, although there is substantial overlap for a subset of pathways.

3. Discussion and Conclusions

The goal of PSB 2016 is to demonstrate advances relative to *"work in databases, algorithms, interfaces, natural language processing, modeling and other computational methods, as applied to biological problems, with emphasis on applications in data-rich areas of molecular biology."* Further *"a major goal of PSB is to create productive interaction among the rather different research cultures of computer science and biology."* The body of work represented by this session, focusing on the development and application of methods for the discovery of molecularly targeted therapies, is emblematic of the vigorous and highly productive exchange of knowledge and ideas surrounding the aforementioned foci. Further, the work summarized herein serves to emphasize:

1) *The state-of-the-art in terms of in-silico knowledge synthesis methods that can be used to identify, aggregate and instantiate component-level models and that can be used to construct application-specific multi-models for therapeutic targeting (e.g., having a specified disease or biological context);*
2) *Ongoing challenges and opportunities surrounding the creation of "interchange layers" and the execution of "vertical reasoning" tasks across and between scalar multi-models in order to generate hypotheses linking synergistic bio-molecular entities or processes of interest and correlative molecularly targeted therapeutic agents; and*
3) *Exemplary instances where the preceding theories and methods have been applied to create an "end to end solution" in which multi-modeling approaches have been used to generate scalar multi-models, identify hypotheses concerning molecularly targeted therapeutics informed by such multi-models, and ultimately evaluate those hypotheses using some combination of in-silico, laboratory, animal or human study paradigms.*

As such, these report amplify the highly promising future for the molecular targeting of therapeutics in a variety of disease states, all in support of what are ultimately envisioned as precision medicine paradigms with the ensuing benefits relative to the quality, safety, outcomes, and costs of such data-driven and adaptive healthcare.

.

References

1. Hood L, Perlmutter RM. The impact of systems approaches on biological problems in drug discovery. Nature Biotechnology. 2004;22(10):1215-7.
2. Auffray C, Chen Z, Hood L. Systems medicine: the future of medical genomics and healthcare. Genome Med. 2009;1(1):2.1-2.11.
3. Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Onc. 2010(8):184-7.
4. Ahn AC, Tewari M, Poon CS, Phillips RS. The Limits of Reductionism in Medicine: Could Systems Biology Offer an Alternative? PLOS Medicine. 2006;3(6):709-13.
5. Jones PA, Baylin SB. The Epigenomics of Cancer. Cell. 2007(128):683-92.
6. Payne PR, Embi PJ, Sen CK. Translational Informatics: Enabling High Throughput Research Paradigms. Physiological Genomics. 2009.
7. Fitzgerald JB, Schoeberl B, Nielsen UB, Sorger PK. Systems biology and combination therapy in the quest for clinical efficacy. Nature Chemical Biology. 2006;2(9):458-66.
8. Barabasi AL, Oltvai ZN. Network Biology: Understanding The Cell's Functional Organization. Nature Reviews Genetics. 2004;5(February):101-13.
9. Anastassiou D. Computation analysis of the synergy among multiple interacting genes. Molecular Systems Biology. 2007;3(83):1-8.
10. Schadt EE, Bjorkegren JL. Network-enabled wisdom in biology, edicine, and health care. Science Translational Medicine. 2012;4(115):115rv1
11. Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med. 2005 May;53(4):192-200.
12. Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. Nature Biotechnology. 2004;22(10):1253-9.
13. Lussier YL, Chen JL. The Emergence of Genome-Based Drug Repositioning. Science Translational Medicine. 2011;3(96):1-3.
14. Ainsworth C. Networking for new drugs. Nature Medicine. 2011(17):1166-8.
15. Blois M. Medicine and the nature of vertical reasoning. N Engl J Med. 1988;381(13):847-51.
16. Tsafnat G, Coiera EW. Computational Reasoning across Multiple Models. J Am Med Inform Assoc. 2009;16(6):768-74.