Network Methods for Multiomic Data Integration

This manuscript (permalink) was automatically generated from zietzm/integration-review@e6daa0c on December 15, 2018.

Authors

- **Michael Zietz**
  0000-0003-0539-630X · zietzm
  Department of Physics, Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania

Abstract

Introduction

In the distant past, medical knowledge advanced toward the goal of finding treatments for every condition, defined very loosely. Once a treatment was found, the mystery disappeared and the disease had essentially been solved. Massive improvements have been made in the quality of medicine and the quantity of medical knowledge under this paradigm. At the current stage, however, diminishing returns can be achieved by simply trying to find a treatment for a broadly-defined disease. Two major avenues for future improvement stand out. First, the implementation of knowledge in real-world clinical settings. Despite the immensity of hard-gotten medical information available, patients continue to die from curable diseases and medical errors. Without putting into practice known ways to improve medicine, future research will fail to achieve its life-saving potential. Second, medical research must broaden its scope to encompass more information while zooming in to identify the specific biological foundations of disease, including the differences between individuals at the molecular level. Research toward the second goal is known as precision medicine, and this review covers a sub-field within precision medicine research.

Omics and precision medicine

Precision medicine entails custom-tailoring health care to individual patients based on their personal, clinical, and molecular profiles. The goal of precision medicine is to improve prevention, diagnosis, and treatment by accounting for the medically-relevant differences between individuals and groups. Toward this goal, the clinical value of omics data and omics-based discoveries has been demonstrated repeatedly. A brief overview of two sub-fields illustrates the possibilities for future research and the potential value of these discoveries.

Pharmacogenomics studies the genomic underpinnings of individual differences in drug response. Compared to the earlier-established field of pharmacogenetics, pharmacogenomics focuses on the simultaneous effects of many genes on drug metabolism and toxicity [1]. As of August 2018, the FDA lists 284 drugs with pharmacogenomic information included on the label [2].

Deep phenotyping attempts to understand and fully classify the biological components underlying disease [3]. This task is made challenging by the fact that there are often not many well-understood links between genotype and phenotype, especially in the context of complex disease. Attention to varied patient information and its meaningful incorporation will require a new paradigm in medical diagnosis. To better understand diverse patient subtypes, consideration must be given to the patient’s full range of other diagnoses and symptoms, even if seemingly orthogonal [4].

The ideal future for precision medicine would allow patients to be categorized into homogeneous groups for prevention, diagnosis, and treatment. That is, precision medicine seeks to uncover and account for the full range of individual variation that underlies differential disease succeptibility, prognosis, and treatment response. Whether in the form of specific clinical traits or detailed multiomic molecular attributes, precision medicine shows promise to overcome the heterogeneity of patients found in previously loosely-defined disease categories. If disease categorizations truly account for the totality of relevant factors, then patients within homogeneous categories can be treated systematically by the establishment of subtype-specific treatment regimes. Working toward such goals requires a broad approach, in which data from many different sources can be incorporated toward a more complete understanding of disease.

Multiomics

While the clinical application of genomic data shows promise, biology cannot be fully explained by a single data type. Many diseases are affected in part by patients' genomic profiles, though few diseases are based soley in genetics. The central dogma of molecular biology states that genetic information flows from nucleic acid to nucleic acid and from nucleic acid to protein. This dogma alone suggests that processes and information beyond the genetic sequence of a patient's DNA will influence phenotype. *Multiomics* refers to the set of such biologically-relevant molecular data sources: genomics, epigenomics, transcriptomics, proteomics, metabolomics, and the organismal microbiome. Complex diseases may be most explainable when viewed through the lenses of multiple such data sources.

In recent years, the cost of omics data generation has fallen dramatically. A former target of $1,000 per genome sequenced now seems high as companies like Illumina promise a future of the, "$100 genome," [5]. Such rapid technological improvements have recently made the creation of large-scale multiomic datasets cost-effective. The Cancer Genome Atlas (TCGA) is a large-scale project which has collected patient information and tumor samples from over 11,000 patients in the United States. Its data include clinical information about patients, sample metadata, and multiomic sample data including gene expression, SNP genotypes, copy number variation, DNA methylation, exon sequencing, and microRNA profiles [6]. Another project underway is the International Cancer Genome Consortium (ICGC), which seeks to incorporate data from TCGA and around 50 other projects to provide researchers access to a massive quantity of tumor omics data [7]. The proliferation and falling costs of next-generation sequencing technologies have allowed these and other large consortia to create massive data resources covering many modalities and diverse diseases. Decreasing costs of data have not, however, resulted in a decrease in the cost of data analysis. Significant future research is needed to deeply probe multiomic datasets in order to extract the data that will guide future diagnostics and treatments.

Data integration

Different data modalities can contain distinct pieces of medically-relevant information. This has led researchers and clinicians to consider how multiple data modalities can together provide a more complete picture of disease. An obvious approach for prediction or classification is the simple combination (or concatenation) of data from multiple modalities. The issue with such approaches has to do with the data themselves. Omics data often contain a very large number of features relative to the number of samples. This leads to a low signal-to-noise ratio, which is only worsened by the concatenation of data from two or more different streams. An alternative and promising approach involves extracting information from data modalities individually before combining extracted data into a single, more cohesive representation with a relatively higher signal-to-noise ratio. Several such techniques have been published and a subset of them will be covered later in this review.

Data-integrative methods are advantageous because they allow researchers to discover connections which are attested by multiple sources of data. This advantage is extremely beneficial, as it allows integrative methods to achieve fewer false positives when compared to single-modality methods. Particularly in the age of high-throughput omics technologies, where data can be staggeringly abundant, false positives are a major problem and steps must very often be taken to reduce their influence.

Another advantage of methods which simultaneously consider multiple sources of data is their superior ability to uncover complex and emergent relationships. Data-integrative approaches to multiomic analysis are better suited to capture nonlinear relationships across data modalities than the simple combination of single-modality-derived conclusions. To better illustrate this point, consider the example biological system proposed by Ritchie et al. [8]. They imagine two competing hypotheses to explain a complex phenotype. First, so-called “Hypothesis A” is a linear system of explanation in which genomic variation leads to variation in gene expression, leading to variation in protein expression, leading to a certain phenotype. The authors propose that Hypothesis A would be best tested by step-wise progress and incremental data reduction/filtering. An alternative system, “Hypothesis B”, the nonlinear hypothesis, involves multi-interacting connections between genome, transcriptome, proteome, epigenome and phenotype. Hypothesis B, the authors suggest, would be better tested using methods that combine the relevant data modalities first, rather than filtering each type to those entities believed *a priori* to be relevant to the phenotype.

As an even simpler example, consider the exclusive or (XOR) function. It is impossible to learn the behavior of XOR while only able to modulate or observe one input. Only by observing both relevant inputs can the function be determined. Similarly, to completely understand complex and nonlinear biological systems, hidden variables must be unveiled and understood in the context of all the relevant data. This task is best achieved by simultaneously considering information from multiple sources, the goal of data integration methods.

Network approaches to data integration

Networks are a logical way of representing many types of biological data. Nodes (or vertices) may represent biological or biomedical entities, while edges may represent a number of different connections or pairings between entities. Several graph-theoretic layers can be helpful in the construction and analysis of biological networks, including graph coloring, edge weighting, and the specification of edge direction. By including such additional layers of information, as well as the association of nodes and edges with biological metadata, networks can consolidate many types of information into a single, cohesive format.

This review focuses on network-based methods for data integration, though such methods are not the only ones which show promise. For further discussion of alternative approaches, several excellent reviews and comparisons are available [10,8,9].

Discussion

Similarity Network Fusion (SNF), a method due to Wang, et al. [11], uses patient similarity networks to cluster patients and predict future labels. In these networks, nodes represent patients and the edges connecting patients represent their pairwise similarity. For each additional data modality, SNF creates a similarity network based on the patients' molecular profiles. The constructed modality-specific networks are then fused using an iterative linear algebra routine which is appropriate for studies of varied sample size and feature number, as well as for highly heterogeneous data sources. The authors applied SNF to three data modalities for a case study of glioblastoma multiforme (GBM), an aggressive brain tumor which is challenging to treat [12]. Previous data-integrative approaches had shown varied results, including the identification of a variable numbers of disease subtypes, depending on the data modality under investigation. Using SNF, the authors were able to recover known clinically-relevant disease subtypes and show that these subtypes correspond to very different survival prognoses. Moreover, by probing the network resulting from the SNF method, the authors were able to show that the majority of the edges were supported by two or more different data modalities.

[13] [14] [15] [16] [17] [18] [19] [20]

[21]

References

1. **Pharmacogenomics: Translating Functional Genomics into Rational Therapeutics**
W. E. Evans
Science (1999-10-15) https://doi.org/cw248b
DOI: 10.1126/science.286.5439.487

2. https://www.fda.gov/drugs/scienceresearch/ucm572698.htm

3. **Deep phenotyping for precision medicine**
Peter N. Robinson
Human Mutation (2012-04-13) https://doi.org/gfpptq
DOI: 10.1002/humu.22080 · PMID: 22504886

4. **Deep phenotyping: The details of disease**
Cathryn M. Delude
Nature (2015-11) https://doi.org/gfpptr
DOI: 10.1038/527s14a · PMID: 26536218

5. **Cheaper DNA sequencing unlocks secrets of rare diseases**
Sarah Neville
Financial Times (2018-03-05) https://ft.com/content/017a3a50-f6f1-11e7-a4c9-bbdefa4f210b

6. **The Cancer Genome Atlas Pan-Cancer analysis project**
John N WeinsteinEric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart
Nature Genetics (2013-10) https://doi.org/f3nt5c
DOI: 10.1038/ng.2764 · PMID: 24071849 · PMCID: PMC3919969

7. **International network of cancer genome projects**
Thomas J. Hudson (Chairperson), Warwick Anderson, Axel Aretz, Anna D. Barker, Cindy Bell, Rosa R. Bernabé, M. K. Bhan, Fabien Calvo, Iiro Eerola, Daniela S. Gerhard, … Huanming Yang
Nature (2010-04-15) https://doi.org/cm9h2m
DOI: 10.1038/nature08987 · PMID: 20393554 · PMCID: PMC2902243

8. **Methods of integrating data to uncover genotype–phenotype interactions**
Marylyn D. Ritchie, Emily R. Holzinger, Ruowang Li, Sarah A. Pendergrass, Dokyoon Kim
Nature Reviews Genetics (2015-01-13) https://doi.org/bg9k
DOI: 10.1038/nrg3868 · PMID: 25582081

9. **More Is Better: Recent Progress in Multi-Omics Data Integration Methods**
Sijia Huang, Kumardeep Chaudhary, Lana X. Garmire

Frontiers in Genetics (2017-06-16) https://doi.org/gcz6m3
DOI: 10.3389/fgene.2017.00084 · PMID: 28670325 · PMCID: PMC5472696

10. **Methods for the integration of multi-omics data: mathematical aspects**
Matteo Bersanelli, Ettore Mosca, Daniel Remondini, Enrico Giampieri, Claudia Sala, Gastone Castellani, Luciano Milanesi
BMC Bioinformatics (2016-01-20) https://doi.org/gcpgct
DOI: 10.1186/s12859-015-0857-9 · PMID: 26821531 · PMCID: PMC4959355

11. **Similarity network fusion for aggregating data types on a genomic scale**
Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, Anna Goldenberg
Nature Methods (2014-01-26) https://doi.org/f5v9f5
DOI: 10.1038/nmeth.2810 · PMID: 24464287

12. https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme

13. **Synergistic effect of different levels of genomic data for cancer clinical outcome prediction**
Dokyoon Kim, Hyunjung Shin, Young Soo Song, Ju Han Kim
Journal of Biomedical Informatics (2012-12) https://doi.org/f4gstk
DOI: 10.1016/j.jbi.2012.07.008 · PMID: 22910106

14. **A Bayesian Integrative Genomic Model for Pathway Analysis of Complex Traits**
Brooke L. Fridley, Steven Lund, Gregory D. Jenkins, Liewei Wang
Genetic Epidemiology (2012-03-28) https://doi.org/f3xkqr
DOI: 10.1002/gepi.21628 · PMID: 22460780 · PMCID: PMC3894829

15. **Time to Recurrence and Survival in Serous Ovarian Tumors Predicted from Integrated Genomic Profiles**
Parminder K. Mankoo, Ronglai Shen, Nikolaus Schultz, Douglas A. Levine, Chris Sander
PLoS ONE (2011-11-03) https://doi.org/c6zcns
DOI: 10.1371/journal.pone.0024709 · PMID: 22073136 · PMCID: PMC3207809

16. **ATHENA: the analysis tool for heritable and environmental network associations**
Emily R. Holzinger, Scott M. Dudek, Alex T. Frase, Sarah A. Pendergrass, Marylyn D. Ritchie
Bioinformatics (2013-10-21) https://doi.org/gfq96v
DOI: 10.1093/bioinformatics/btt572 · PMID: 24149050 · PMCID: PMC3933870

17. **A statistical framework for genomic data fusion**
G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, W. S. Noble

Bioinformatics (2004-05-06) https://doi.org/bsph7z
DOI: 10.1093/bioinformatics/bth294 · PMID: 15130933

18. **Protein function prediction via graph kernels**
K. M. Borgwardt, C. S. Ong, S. Schonauer, S. V. N. Vishwanathan, A. J. Smola, H.-P. Kriegel
Bioinformatics (2005-06-01) https://doi.org/c6s8jt
DOI: 10.1093/bioinformatics/bti1007 · PMID: 15961493

19. **Fast protein classification with multiple networks**
K. Tsuda, H. Shin, B. Scholkopf
Bioinformatics (2005-09-01) https://doi.org/bhdtqm
DOI: 10.1093/bioinformatics/bti1110 · PMID: 16204126

20. **Graph sharpening plus graph integration: a synergy that improves protein functional classification**
Hyunjung Shin, Andreas Martin Lisewski, Olivier Lichtarge
Bioinformatics (2007-10-31) https://doi.org/dt399h
DOI: 10.1093/bioinformatics/btm511 · PMID: 17977886

21. **The properties of high-dimensional data spaces: implications for exploring gene and protein expression data**
Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, Yue Wang
Nature Reviews Cancer (2008-01) https://doi.org/ffksnf
DOI: 10.1038/nrc2294 · PMID: 18097463 · PMCID: PMC2238676