Network Methods for Multiomic Data Integration

This manuscript (permalink) was automatically generated from zietzm/integration-review@6c6f488 on December 19, 2018.

Authors

- **Michael Zietz**
  0000-0003-0539-630X · zietzm

  Department of Physics, Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania

Abstract

The age of abundant data has arrived. As the costs to generate and distribute omics-scale data have plummeted in the last two decades, previously impossible analysis is now reality. Through the integration of many types of omics data, diseases can increasingly be understood on their own terms: as highly complex, self-interacting, nonlinear biological systems. This review seeks to describe some of the many promising methods to simultaneously leverage multiple data modalities. Specifically, we examine the application of network methods as powerful, extensible, and natural ways to deal with multiple streams of information about a large number of entities. Significant progress has been made in the past two decades in unsupervised methods for data integration. While unsupervised methods can be extremely beneficial to supervised analyses, such applications remain an area of active and vibrant research. We conclude the review by exploring some of the potential avenues for future research and discovery.

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 susceptibility, 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 (CNV), 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 [10,10,8,9].

Discussion

Machine learning techniques have broadly been classified into two disjoint categories, supervised and unsupervised.[1] While in reality the two categories represent endpoints of a continuum, this review will consider the only the binary discretum. Through such a lens, “supervised” refers to techniques which attempt to learn mappings between specified inputs and outputs. A simple example of a supervised learning technique is logistic regression, which attempts to assign binary labels to data points after being trained on labeled data. “Unsupervised” learning refers to techniques which identify characteristics of data without labels. A simple example of unsupervised learning is principal component analysis (PCA), which transforms observations onto a new set of coordinates that represent the axes of greatest variance. The discussion that follows is split between supervised and unsupervised methods. At present, greater advances appear to have been made in unsupervised methods, though supervised methods can be strengthened and enabled by first applying unsupervised data integration, feature selection, or de-noising.

Unsupervised methods

A very useful application of unsupervised methods is the transformation of large noisy data into a useful, de-noised format which can used for supervised machine learning. For example, Greene et al. [11] constructed tissue-specific functional gene networks using data from over 14,000 experiments. By integrating tissue-specific expression information in a Bayesian framework, the authors constructed networks of tissue-specific gene functional relationships, with edge weights representing the posterior probability of a functional relationship. These edges weights were determined by a naive Bayes classifier that was trained to predict positive functional relationships based on tissue-specific data. A key insight by the authors was that tissue-specific networks represent a data-driven method for the analysis of noisy GWAS data. The corresponding method is outlined later in this review.

Another data-integrative application of functional networks was proposed by Sehgal et al. [12], who developed a methodology called the Robust Selection Algorithm (RSA) to identify micro-RNAs (miRNA) which are oncogenic or tumor-suppressing. By incorporating functional networks of miRNA-mRNA data, the authors were able to account for non-linear relationships between molecular markers and patient outcomes. Using data from TCGA, RSA was able to identify a number of both onco-miRNA and tumor-suppressor miRNA with relevance for the prediction of survival times in various cancers. To generate functional networks, the authors employed the NetWalker [13] suite of software, a self-described, “one-stop platform,” for network-based data methods for integrative genomics.

A common goal of network-based systems biology is the identification of functionally-related modules (or sub-networks) within larger networks of genes or proteins. Ideker et al. [14] developed an approach to integrate protein-protein and protein-DNA interactions with mRNA expression data to identify “active” subnetworks. Defined as, “connected sets of genes with unexpectedly high levels of differential expression,” active subnetworks were identified by finding groups of nodes with significant differential expression, the *collective* expression of which was compared to randomly chosen subnetworks. Having scored subnetworks, the authors’ method then finds the highest-scoring subnetworks using simulated annealing. Subsequent research in module network inference has expanded the methods’ capabilities to most common types of omic data. Lemon-Tree, due to Bonnet et al. [15], is an open-source software package that encompasses many validated algorithms and novel approaches for module network inference.

iCluster [16] is a joint latent variable model, which attempts to estimate unobserved tumor subtypes using available multiomic data such as CNV, DNA methylation, and mRNA expression. The method estimates tumor subtypes by maximizing a set of linear equations using maximum likelihood and the expectation-maximization algorithm. Applying the method to breast cancer, the authors were able to identify four disease subtypes, recovering well-known clinical subtypes, as well as identifying potentially novel subtypes of the disease.

Network-based methods for biomedical data integration are not limited to representations in which nodes represent genes, proteins, RNA molecules, or other molecular entities. Similarity Network Fusion (SNF), a method due to Wang, et al. [17], uses patient similarity networks to cluster patients into discrete and clinically-meaningful groups. 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 [18]. 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, strengthening confidence in the validity of their integrative method.

Applying individual weights to networks is a natural next-step for the fusion of multiomic networks. In a separate application of SNF, Angione et al. [19] developed a network approach for bacterial growth and response, in which nodes represent growth conditions and edges represent the similarities between conditions. In a modification to the original SNF method, Angione et al. propose a weighted or "biased" SNF method in which layers themselves can be weighted according to the predictive quality of the various data. The authors state that this modification also increases the method's robustness to distributions with high kurtosis. After consolidating the network into a single, global representation of the multi-modal data, the authors were able to predict *E. coli* metabolism under a range of possible growth conditions.

Mosca et al. [20] developed a multi-objective data fusion method, in which both component- and network-level estimators are optimized. As in the biased SNF method, networks themselves are weighted during the integration procedure. Unlike the previous work, though, the multi-objective method finds optimal weights for both networks and nodes, thus leaving fewer hyperparameters needing user-given specification. A difficulty of the MO method lies in the definition of objective functions to be minimized. While the authors propose several (over-representation analysis statistics, functional class scoring à la GSEA, etc.), it is not clear which metric would be most appropriate for any given situation. Nonetheless, the method presented is an interesting and novel way to optimize the multiomic sets of networks that arise in integrative genomics.

Another key application for unsupervised data integration methods is the prioritization or ranking of the molecular causes of disease. Greene et al. [11] developed a method called NetWAS to re-prioritize genes with nominally-significant p-values from standard GWAS methods. NetWAS uses tissue-specific functional networks as features when training a support vector machine (SVM) classifier to predict whether a gene was nominally significant in a GWAS study. After training, genes can be re-ranked according to their distance from the SVM decision hyperplane. The rationale behind this method is that the SVM is able to capture some characteristics of the data which lead to significant gene enrichment, while genes with spurious p-values will not share the same functional characteristics. One of the pivotal innovations of NetWAS is the ability to re-prioritize or cross-prioritize genes from distinct GWAS studies in a method that does not depend on *a priori* knowledge of the disease.

Aerts et al. [21] also developed a method to prioritize the genes that potentially underlie disease. Named Endeavor, the method ranks genes based on their similarity to genes known to be involved in the disease. After producing a ranking using each data modality available separately, the ranking lists are combined. In a later paper, De Bie et al. [22] improved upon the Endeavor methodology by devising a kernel-based method to rank genes within a single data modality. Both of these methods have the disadvantage of needing comparison data in which to contextualize unseen observations.

Overall, unsupervised data integration methods have made tremendous progress in the past two decades. These methods have shown promise for the generation of useful datasets, the data-driven identification of disease subtypes and patient clusters, as well as for the extraction of information pertinent to prediction tasks.

Supervised methods

While supervised learning methods are methodologically quite distinct from unsupervised approaches, the underlying learning is fundamentally similar. In both cases, the goal is to develop computational models that are able to learn about the intrinsic qualities, connections, and variation patterns within high-dimensional, noisy data. Supervised learning then leverages this lower-level understanding to make explicit and testable predictions about observed data.

A classic application of supervised learning to networks is the task of edge prediction. Within networks of a single type, several common methods exist. A basic approach involves dropping several nodes from the network, computing walk-counts between pairs of nodes, then using these counts as features to predict the edges which were dropped. Superior methods have been devised, but few are immediately applicable to multi-modality datasets in a way that would produce meaningful and interpretable results. Another key application of supervised learning is the prediction of node sets that are related in some way or relevant to some other phenomenon. Among the most common applications is the prediction of genes that are relevant to a particular disease.

Himmelstein et al. [23] created a so-called “heterogeneous” network with nodes representing biomedical entities of many different types. Fundamentally, the authors’ data integration strategy is simply the joining of multiple types of relationship data (eg: genes, diseases, molecular functions, compounds) along sets of common vocabularies. After creating a multipartite network with over 40,000 nodes and 1,600,000 edges, the authors performed a number of prediction tasks, illustrating the power of integrative approaches. As features to their predictive models, the authors devised a novel metric of node pair interconnected-ness, termed the "degree-weighted path count," (DWPC). Computed by traversing the network across multiple semantic types of paths (eg: Gene-Associates-Disease-treated by-Compound), DWPC allowed the authors to predict withheld high-confidence associations from the GWAS catalog at an area under the receiver operating curve (AUROC) of 0.83.

In a follow-up application [24] the authors utilized a similar method to predict candidates for drug repurposing within a much-enlarged network. Predicting treatments for epilepsy, the authors were able to recapture 23 of 25 (withheld) known anti-epileptic drugs. One of the advantages of the heterogeneous network method is that connections between any two types of biological entities in the network can be predicted using the same methodology. A drawback of this method is that gold-standard data must be obtained for the connection-type to be predicted, which could be challenging for certain types of biological connections.

Hypothesizing that many genes involved in cancer development may not be identified by conventional methods such as differential expression or mutation analysis, Ruffalo et al. [25] devised a method to identify these so-called, "Silent players." Fundamentally, the method seeks to identify proteins involved in cancer by computing the proximity of all human proteins to products of genes which are differentially-expressed or mutated. "Proximity" in this context is computed by network propagation and results in two vectors (one for each of the input data modalities). These features are then used to predict the probability of the gene being involved in the disease of interest. While the method described in the paper integrates only mutation and differential expression data with protein-protein interaction networks, the method could, in principle, be applied to similar data modalities as well, provided suitable networks exist for the modalities of interest.

Supervised methods also lend themselves to the prediction of clinical outcomes. Kim et al. [26] used multiomic data to predict outcomes for several cancers. Their data integration method first created a patient similarity network for each data modality. Next, the individual networks were merged by finding the optimal coefficients for a linear combination of graphs such that an objective function involving the graph Laplacian was minimized.[2] This work preceded the publication of SNF, though it is similar in approach and in effect. After merging modality-specific networks into a global network of patient similarities, the authors predicted clinical outcomes using a nearest-neighbor approach.

Shin et al. [27] created an integrated network of proteins by finding the optimal linear combination of individual graph Laplacians. Importantly, they developed a method called graph sharpening, in which edges between labeled and unlabeled points are removed from the graph. Sharpening reduces the noise found in a graph by eliminating undesired edges. The authors found that this method, when applied in conjunction with network integration, affords the greatest improvement for the task of predicting whether proteins are members of a gene ontology (GO) category.

Other clinical events beyond the emergence of disease can be predicted as well. Mankoo et al. [28] developed a model with a focus on reducing the number of features used for prediction and increasing the biological interpretablility of outcomes. In the study, the authors sought to predict clinical outcomes for patients with serous ovarian cancer using multiomic data from TCGA. Their method filtered features by computing the Spearman rank correlations between data types and removing all features below chosen cut-offs. Overall, this procedure reduced the number of features from roughly 50,000 to under 200. Using the reduced feature set, the authors predicted discrete outcomes and continuous time to recurrence better than with any of the data types alone.

Conclusion

In this review, we have described many of the major recent advances in network-based methods for multiomic data integration. Networks are natural representations of many types of biological and biomedical data, and they can be very effectively leveraged for integrative analyses.

Unsupervised methods attempt to integrate data for the purposes of feature reduction, de-noising, clustering, stratification, and future prediction methods. Another application which is particularly relevant to unsupervised network methods is the identification of functional modules (or sub-networks) that correspond to biologically-meaningful functional units. Significant conceptual advances include the creation of networks in which nodes have unconventional meaning, such as bacterial growth conditions or patients themselves. Unsupervised integrative analyses have also shown their power for the prioritization (or re-prioritization) of molecular entities according to their potential influence on phenotype.

An intriguing future direction of research involves the application of novel network methods to integrated omics networks. For example, Tan et al. [29] used an ensemble of unsupervised methods to extract molecular signatures from gene expression data. The authors employed an ensemble of denoising autoencoders to extract stable molecular signatures from the highly over-parametrized model. Similar ensemble-of-unsupervised approaches could be applied to integrate multi-modal omics data.

Supervised methods attempt to harness multifarious data to make biological or clinical predictions. While supervised methods lend themselves less naturally to the practice of data integration, many unsupervised methods return integrated data which is particularly amenable to supervised prediction. A major future advancement in this field will undoubtedly involve the association of outcome labels with multiomics data of a massive scale. Specifically, the integration of omics data with electronic medical records (EMR) could lead to the creation and availability of huge data sets alongside information about disease, treatment, and response. Such an integration is one of the most promising avenues for future work in this field, and it would lay the path for significant future discovery. Among the many types of research that such integration would allow, empirically derived “deep phenotypes” are some of the most exciting possibilities.

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. **Table of Pharmacogenomic Biomarkers in Drug Labeling**
U.S. Food and Drug Administration
U.S. Food and Drug Administration (2018-08-03) 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. **Understanding multicellular function and disease with human tissue-specific networks**
Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, Ran Zhang, Boris M Hartmann, Elena Zaslavsky, Stuart C Sealfon, … Olga G Troyanskaya
Nature Genetics (2015-04-27) https://doi.org/f7dvkv
DOI: 10.1038/ng.3259 · PMID: 25915600 · PMCID: PMC4828725

12. **Robust Selection Algorithm (RSA) for Multi-Omic Biomarker Discovery; Integration with Functional Network Analysis to Identify miRNA Regulated Pathways in Multiple Cancers**
Vasudha Sehgal, Elena G. Seviour, Tyler J. Moss, Gordon B. Mills, Robert Azencott, Prahlad T. Ram
PLOS ONE (2015-10-27) https://doi.org/f78knm
DOI: 10.1371/journal.pone.0140072 · PMID: 26505200 · PMCID: PMC4623517

13. **NetWalker: a contextual network analysis tool for functional genomics**
Kakajan Komurov, Serkan Dursun, Serkan Erdin, Prahlad T Ram
BMC Genomics (2012) https://doi.org/gb3fdc
DOI: 10.1186/1471-2164-13-282 · PMID: 22732065 · PMCID: PMC3439272

14. **Discovering regulatory and signalling circuits in molecular interaction networks**
T. Ideker, O. Ozier, B. Schwikowski, A. F. Siegel
Bioinformatics (2002-07-01) https://doi.org/cwkstn
DOI: 10.1093/bioinformatics/18.suppl_1.s233

15. **Integrative Multi-omics Module Network Inference with Lemon-Tree**
Eric Bonnet, Laurence Calzone, Tom Michoel
PLOS Computational Biology (2015-02-13) https://doi.org/gcpgc3
DOI: 10.1371/journal.pcbi.1003983 · PMID: 25679508 · PMCID: PMC4332478

16. **Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis**

Ronglai Shen, Adam B. Olshen, Marc Ladanyi
Bioinformatics (2009-09-16) https://doi.org/dvdz5s
DOI: 10.1093/bioinformatics/btp543 · PMID: 19759197 · PMCID: PMC2800366

17. **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

18. **Glioblastoma Multiforme**
American Association of Neurological Surgeons
American Association of Neurological Surgeons (2018-08-03) https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme

19. **Multiplex methods provide effective integration of multi-omic data in genome-scale models**
Claudio Angione, Max Conway, Pietro Lió
BMC Bioinformatics (2016-02) https://doi.org/gfq976
DOI: 10.1186/s12859-016-0912-1 · PMID: 26961692 · PMCID: PMC4896256

20. **Network-based analysis of omics with multi-objective optimization**
Ettore Mosca, Luciano Milanesi
Molecular BioSystems (2013) https://doi.org/gfq996
DOI: 10.1039/c3mb70327d · PMID: 24121459

21. **Gene prioritization through genomic data fusion**
Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, … Yves Moreau
Nature Biotechnology (2006-05) https://doi.org/bcmjgf
DOI: 10.1038/nbt1203 · PMID: 16680138

22. **Kernel-based data fusion for gene prioritization**
Tijl De Bie, Léon-Charles Tranchevent, Liesbeth M. M. van Oeffelen, Yves Moreau
Bioinformatics (2007-07-01) https://doi.org/b46cvm
DOI: 10.1093/bioinformatics/btm187 · PMID: 17646288

23. **Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes**
Daniel S. Himmelstein, Sergio E. Baranzini
PLOS Computational Biology (2015-07-09) https://doi.org/98q
DOI: 10.1371/journal.pcbi.1004259 · PMID: 26158728 · PMCID: PMC4497619

24. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

25. **Network-Based Integration of Disparate Omic Data To Identify “Silent Players” in Cancer**
Matthew Ruffalo, Mehmet Koyutürk, Roded Sharan
PLOS Computational Biology (2015-12-18) https://doi.org/gfq992
DOI: 10.1371/journal.pcbi.1004595 · PMID: 26683094 · PMCID: PMC4684294

26. **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

27. **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

28. **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

29. **Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks**
Jie Tan, Georgia Doing, Kimberley A. Lewis, Courtney E. Price, Kathleen M. Chen, Kyle C. Cady, Barret Perchuk, Michael T. Laub, Deborah A. Hogan, Casey S. Greene
Cell Systems (2017-07) https://doi.org/gcw9f4
DOI: 10.1016/j.cels.2017.06.003 · PMID: 28711280 · PMCID: PMC5532071

[1] An oft-included third category is reinforcement learning, in which a computational agent learns actions to maximize a reward function. While such methods show promise for biological applications like the prediction of protein folding, they are beyond the scope of this review.↩

[2] Note, the Laplacian for a graph is defined as the diagonal degree matrix (elements on the main diagonal correspond to the degree of the respective node) minus the adjacency matrix.↩