Computing with Interaction Networks (Biomedical Data at the Molecular Scale)

0. ABSTRACT (WUM):

Interactions between biomolecules are at the core of human biology. Disease arises not only from single molecular defects but also from disturbed interactions between many proteins and functional genomic elements. The same interactions that make life so complex and wonderful, also make some diseases difficult to treat. Network theory is a well-developed branch of mathematics that organizes and analyzes the interactions of parts within a system. Network theory is of particular relevance to biology and medicine, as it provides tools and a framework for understanding molecular interactions. Through a network approach to biomedical data, insights from across diverse fields can be brought to bear on biomedical data.

0.*1 Keywords:*

Network analysis, molecular interaction, systems biology, cross-disciplinary research, network medicine.

1. INTRODUCTION:

1.1 Networked systems are at the core of human biology (PDM):

Human health and disease depend on vast networks of molecules that interact and communicate. Transfer of genetic information, cellular communication, and human metabolism are all mediated by complex pathways and networks of molecules. Network analysis of large-scale data has been used to identify critical pathways and proteins in gene regulatory networks {24092746}, including molecular pathways affected by cancer \{22955619, ENCODE and Cancer?}. Off-target effects of prescription drugs have been predicted through a network model of metabolism \{23455439}. Insights into inflammatory diseases like asthma have been revealed by studying the structure and function of networks of inflammatory signaling molecules \{23407534, 25981665, 17962519}.

Molecular networks change and evolve over time with surprising dynamic complexity \{15372033}. Pro-inflammatory T-cells of the immune system rewire their regulatory networks in autoimmune disease \{23467089, 27307629}. The microbiome of the gut interacts with the human metabolome, and both change together in response to diabetes, or pregnancy, or antibiotic treatment \{22863002, 26633628,24445449}. Substantial changes in the epigenome are observed in human tissues according to cell-type \{25693563}. Network rewiring may be both the cause and consequence of changes to human health \{19741703}. Complete understanding of many molecular networks requires an understanding of these temporal features.

The temporal evolution of molecular networks allows them to perform logical operations and transmit complex signals \{14530388}. Exciting discoveries have been made related to the possibility of logic based communication on networks. For example, this network logic appears to be critical during the orchestration of embryonic development \{23412653,22927416}. There is a tantalizing possibility for future bioengineering of molecular interaction networks to perform

complex logic, and to intervene in disease processes {23041931, 24908100}. A greater understanding of biological networks and their logical structures may eventually provide a platform for augmentation of existing biological capability.

There is potential for theory and techniques from across the well-developed field of network-science to be gainfully applied to molecular data. Novel network analysis techniques like HotNet \{25501392,21385051} use the principle of 'guilt by association' to identify associated molecular function in molecular networks. Advanced machine learning techniques like the deep neural network method DeepBind, have resulted in the discovery of network topology through large-scale analysis of genetic sequences \{26213851}. Connectedness among molecular structures means that network-based techniques are a natural fit for analyzing large data sets of molecular information \{19741703}.

Network analysis of biomedical data is not just a research technique but is critical to the practice of modern medicine. Autism and schizophrenia are among diseases that we now understand are unlikely to be associated with a single molecular alteration, but by multiple affected genes in critical molecular pathways\{27479844, 23453885}. Clinical use of gene expression panels, like the 21-gene panel OncoType Dx that predicts breast cancer recurrence, use molecular phenotype as a proxy for disease phenotype \{26412349}. Disease transmission through social networks, as in the 2013 ebola-virus outbreak in West Africa \{26465384} or Zika-virus spread in the Americas \{28538723, 27013429}, may be tracked through molecular signatures left by the virus as it spreads. These examples suggest the value of the application of network analysis techniques to medicine.

Insight into biological networks may be gained through cross-disciplinary network comparisons. It appears the gene-regulatory network of *E. Coli* is built for robust function in comparison with a computer network that prioritizes economy and reuse above redundancy \{cite}. Using the context of a social network, apparently distant connections between immune cells, may, in reality be closer than appearance \{28263321}, and 'cross-talk' between immune cells may modulate the body's immune response \{24923297}. Evidently by comparing networks from different disciplines through analogy, we may gain insight into both networks.

1.2 Networks leverage abundant biomedical data (PDM):

The Human Genome Project arguably represents the first big-data and large-scale science project in biology \{12690187}, and marked the transition of molecular biology from a 'data-poor' to a 'data-rich' field \{12432964}. It was this transition that was a motivator for the development of the discipline of systems biology \{12432964, 20604711}. When large-scale science projects that produce 'parts list' of molecular structures and entities, systems biology seeks to understand how these parts are connected. Network theory became a foundational technique making sense of these increasingly large data sets of connected biomolecules.

Molecular biology projects continue to expand in size and scope. Genome-scale network reconstructions of metabolic networks have been produced for hundreds of species, and are constantly undergoing refinement \{24811519, 27893703}. The recently released BioPlex 2.0 is the largest protein-protein interaction ever built, with 56,000 listed interactions \{28514442}. Whole genome sequencing projects like 100,000 Genomes project, and the NIH's Genome Sequencing Project, now seek to enroll hundreds of thousands of participants \{26310768,

https://www.genome.gov/27563453/}. Visions have been presented for even larger scale sequencing \{26430149, 23138308}, and the growth of big data in genomics may outpace big data growth in other data intensive fields \{26151137}.

Networks produced from data of this scale have been likened to a 'hairball' when visualized, suggesting their complexity \{27047991}. Identifying meaningful structure and function in these hairballs represents a challenge in the field of biology. The application and development of computational network approaches represents one of the most promising means of unravelling the complicated patterns of connection in these networks \{27387938, 27387949, 23194371}.

The importance of network techniques for analyzing large-scale molecular interaction data is further stressed by the need to integrate diverse sources of molecular data. The number of advanced functional molecular assays available to researchers continues to grow through projects like ENCODE \{25693563}, and new network-based approaches for integrating large-scale biological data are being developed \{24464287, ENCODEC?}. Integration of functional genomics data has been proposed as the clearest way-forward to understanding the significance of human genetic variation \{20020535, Functional precision cancer medicine—moving beyond pure genomics}. Network approaches play a central role in the integration of these diverse sources of large-scale molecular interaction data.

2. MODELING A MOLECULAR INTERACTION NETWORK

2.1 Basic features of an abstract molecular interaction network (PDM):

Before discussion of more advanced techniques for modelling and analyzing molecular interaction networks, we'll present a few widely used definitions and principles that serve as building blocks for more advanced methods.

Central to an interaction network, is a collection of biomolecules with evidence of direct interface of their molecular surfaces [Figure 1.a.1]. This is a 'parts list' of molecular entities, without labeled connections. If the pattern of connections between molecules is known, a network can be formed [Figure 1.a.2]. Upon such a basic network, a progressive layering of information and logic can be tailored according to the network under study. For example, the direction of connections [Figure 5.a.3] and the weight of connections [Figure 1.a.4] may be important information for regulatory networks and gene co-expression networks, respectively.

Higher order relationships between molecular species are also possible. Arbitrarily complex computation can be performed on a network, and abstracted in the form of logic modules or motifs [Figure 1.a.5]. Molecular networks and logic performed by the network exist in 3 spatial dimensions, and this 3D spatial information can be important to understanding the structure and function of a molecular interaction network [Figure 1.a.6].

Matrix representations of interaction network variables are also possible for some networks. Matrix representation of the connections, weight, and direction of connections in hypothetical interaction networks are shown in Figure 1.c.

This set of network variables (connections, direction, weight, time-dependent logic, and spatial geometry) are basic building blocks that network scientists use to describe molecular interaction networks. In addition to these basic building blocks, summarized in Figure 1, a pictorial glossary of network terminology is presented in Figure 2.

2.2 *Incorporating molecular structure in a network model (SK):*

Although there are advantages to abstract representations of molecular networks, there are also inherent limitations. For instance, protein-protein interactions are often represented as a network. Nodes in this network correspond to individual proteins and edges represent interactions between them. Such abstract representations are helpful to gain insight into the overall topological properties of the network. Furthermore, one can identify key proteins based on their connectivity in the network. However, such abstract representations do not provide any biophysical insight into interactions underlying protein-protein interactions.

To address this issue, various studies have integrated three-dimensional structural information data available for various biomolecules to produce structural interaction network (SINs) \cite{17185604,18364713, 21826754} [Figure 3]. Integration of structural information can help address key issues. For example, one can identify key residues or domains on the surface of proteins, which are involved in interactions. In addition, structural information is helpful to predict binding affinities and kinetic constants of the underlying interactions. Furthermore, SINs are helpful in identifying obligate (permanent) or transient interactions in a network. Structural information can also help to distinguish between simultaneous and exclusive interactions. These are key network properties, which cannot be addressed with a simple abstract representation of the network. Finally, integration of structural information can help in gaining mechanistic understanding of rare or disease-associated mutation impact on protein interactions \cite{27915290}. Structural interaction networks can thus be used to prioritize variants in a disease cohort or rare deleterious variants in a population level study.

2.3 Network 'rewiring' - the time based evolution of molecular networks (DL):

Biological networks are hardly static; they may evolve slowly over time or transform rapidly in order to adapt to an environmental change, throughout the development \{20486137}, or simply, as a result of accumulation of mutations. In the context of biological networks, rewiring refers to a complex reformation of interacting partners, such as genes, proteins, and other biologically relevant chemicals [Figure 4].

The central concepts of network rewiring have been around for decades. There have been previous attempts to understand the network dynamics by comparing transcription factor-gene networks in different conditions, but the scope of these efforts were limited to availability of data \{15372033}. The advent of large-scale genomic and proteomic surveys allowed for creation of different types of biological networks, including protein-protein interaction networks (PPIs) and gene regulatory network (GRNs), in different cellular contexts. While it is still difficult to grasp the dynamic nature of biological networks, these advanced assays can provide clearer insight into how genes and proteins operate in point-in-time snapshot, and researchers have been trying to stitch these snapshots back to answer more complex questions in systems biology.

Many studies have focused on the broadest timescale for network rewiring by linking the evolutionary changes of biological networks to diversity among species \{26657905}. In specific, it has been shown that regulatory changes in transcription factor-target networks may account for the species differentiation \{17690298, 20378774, 21253555, 23198090}. However, researchers have also attempted to interpret the network rewiring at much shorter timescales. It is possible to introduce an artificial perturbation into a network and examine the rewiring consequences. One study on a bacterial gene network has shown that perturbations that span four orders of magnitude can propagate and alter up to approximately 70% of the transcriptome \{26670742}. Rewiring has often linked as a result of mutations. A single mutation placed at a regulatory protein binding site can alter the binding specificity, perturbing its interacting neighbors, and consequently, it could have a detrimental downstream effect on the whole network.

Naturally, many studies have attempted to measure the rewiring to infer the consequence for disease phenotype. For example, cancer mutations could affect both downstream and upstream rewiring of the regulatory network, altering cell-signaling and regular gene expressions \{26388441,26388442}. To measure the rewiring, target changing, of TF-gene network involves comparing of networks at two states, before and after the rewiring. Regulatory interconnection between genes can be represented as ones that are gained, lost, or retained. As a result, network rewiring can change gene hierarchy. One study showed rewiring of gene network can promote or demote the importance of a gene as regulator \{21045205}.

More recently, CRISPR genome editing technology has been developed and widely applied in the field of genomics, allowing us to build more complex models to test the effects of cancer mutations. It could prove to be an excellent tool to experimentally validate the results of rewiring obtained via an integrative approach.

Link

2.4 Network motifs, network logic, and network stability (MTG):

Most biological networks, such as protein-protein interaction network, have evolved to maximize network efficiency, functionality, and stability. From this standpoint, to fully reveal the underlying mechanisms of the biological networks that we study, it is important to understand the organizing principles of biological network structure. Network structure evolves alongside biological function, and lays the foundation for complex network processes.

Studies have shown that small structurally stable network motifs are enriched in transcription regulatory networks and perform various functions \{16187794}. Negative autoregulation motifs, for example, have been shown to shorten the response time of stimuli-induced gene expression regulation, as well as reduce cell-cell variation in protein levels. Feedforward loops are another frequently observed motif in gene regulation networks [Figure 5]. Feedforward loops can filter persistent signals from brief spurious pulses of signal, and are also capable of generating pulses and accelerating biological responses. Combinations of network motifs allows for more precise control of biological systems, including the temporal order of gene expression and oscillations in expression \{17510665}.

How

Biological networks have also developed structure to enhance stability. The molecular network, for example, is subjected to exogenous attacks or endogenous mutations that result in dysfunction. A cascading deleterious effect could propagate via links in the network. An

observed feature of many molecular interaction networks is the duplication of extremely vital hubs. Multiple and repeated domains are enriched in hub proteins. \{16780599}. While redundancy may lead to inefficiency, biological networks must balance between stability and energy-loss.

3. TOOLS AND ALGORITHMS FOR NETWORK ANALYSIS

3.1 Advances in network algorithms -- network propagation methods (DC):

In biology and other disciplines, networks have long been used to study complex associations within large datasets. In the context of biology, such datasets include physical interactions between proteins (i.e., protein-protein interaction networks), regulatory relationships (e.g., associations between transcription factors and target genes or miRNAs and their associated targets), or directed pathways of interacting cellular species. As these datasets grow in size, the associated networks used to describe them grow in topological complexity. Positively identifying true signals in these networks can be difficult to attain, given the noise and complexity that accompany the large detests. Recently developed algorithmic frameworks have been developed to capture difficult-to-discern relationships between genes, as well as to identify sub-networks that may be dysregulated. Along these lines, algorithms based on network propagation have proven to be the most powerful \{28607512} [Figure 6.a.].

Generally speaking, the term “network propagation” refers to the analysis of networks by allowing some form of information to flow from node another via shared edges \{26683094, 22035267}. This information may traverse from node to node as a random walk, for instance. Edges may also be weighted (by confidence of an interaction, for example) to influence the “current” of information traveling from one node to another.

Other approaches at inferring gene-gene associations include direct neighbors or shortest paths. Such methods may suffer from high rates of false positives or false negatives, whereas propagation-based methods may optimally capture known gene-gene associations. For instance, Ruffalo et al. use propagation to positively identify cancer-associated genes using both somatic variant data and gene expression as input to the original network \{26683094}.

Such methods have also been leveraged to identify cancer sub-types based on patient stratification \{24037242}, and they have also been used in an array of other disease contexts \{26963104, 27307626, 27489002, 24464287}.

3.1 Advances in network algorithms -- machine learning and neural networks (Holly Zhou):

Machine learning, especially deep learning, is valuable for networks analysis because its multilayered neural networks can “learn” complex patterns and multi-level information processing within cells. The learning layers can reduce complex, many-dimensional data into lower-dimensions at each layer and integrate diverse data types at higher layers \{26252139}, thereby making complex networks more tractable to regulatory genomics studies. By finding hidden patterns in large datasets, machine learning models can predict relationships in networks without requiring strong assumptions about underlying mechanisms \{27474269}. A machine learning workflow involves gathering vast amounts of data, preprocessing the data,

training and testing a model, and interpreting results. At each layer of the neural network, inputs are transformed by a nonlinear processing unit (activation function) and fed into the next layer. The processing unit performs feature extraction, which extracts information to pass into the next layer to facilitate learning.

We use two example problems to motivate our coverage of popular machine learning techniques, as well as to highlight advantages and disadvantages of current methods: 1) predicting the effects of noncoding variants, and 2) predicting transcription factor (TF) binding sites.

Eukaryotes have complex regulatory networks--we still do not know the mechanism of many noncoding variants in genomic data, some of which are linked to human disease. Given a single sequencing assay data input, Basset can learn a cell's chromatin accessibility code to better annotate genome mutations \{27197224}. A primary motivation for this is to link genomic variants in poorly-annotated noncoding regions to phenotypes \{19474294}. Unlike standard neural nets, Basset uses a convolutional neural networks (CNN) with three convolution layers and two fully connected layers [Figure 6.b.]. In standard neural nets (fully connected networks), each hidden layer contains independent neurons that receive a vector input; layers are fully inter-connected. The last layer, which is also fully connected, outputs the prediction (e.g., classification score). CNNs allow for greater complexity--they usually contain an input layer, convolution layer, rectified linear unit (ReLU) layer, pool layer, and a fully-connected layer (see FIGURE for an explanation of CNNs). CNNs allow researchers to study SNVs with much greater resolution and to learn more abstract representations of the regulatory DNA--researchers can prioritize mutations predicted to contribute to regulatory activity and deprioritize others \{27197224}.

Notable methods for modeling DNA and RNA targets of regulatory proteins include DeepBind \{26213851}, DeepMotif \{https://arxiv.org/abs/1605.01133}, and a deep-learning based imputation method for TF binding predictions \{28234893}. Identifying these targets is important for modeling biological processes and for discovering variants that could cause human disease.

DeepBind uses the aforementioned convolutional network architecture with mini-batch stochastic gradient descent to predict the sequence specificities of DNA- and RNA-binding proteins. Mini-batch stochastic gradient descent (SGD) seeks to minimize an objective function; while the standard gradient descent tries to converge for each training example, mini-batch SGD computes the gradient against more than one sample at each step for a smoother convergence. To prevent overfitting, the model uses dropout regularization, weight decay, and early stopping \{26213851}. Dropout regularization refers to randomly setting neurons to have a value of 0 in the intermediate steps, or "dropping" them. Weight decay penalizes large weights—this improves the model's generalization. DeepBind improves upon prior motif-scanning algorithms by taking into account RNA-binding proteins that can recognize secondary or tertiary structural elements. It also recognizes higher-order structures that result from competitive or synergistic effects of protein binding \{26213851}. DeepBind takes as input data from high throughput experiments and is validated through Pearson correlation, Spearman correlation, and AUROC comparisons with 26 published algorithms.

DeepMotif improves on the TF binding classification of DeepBind through a deep CNN and a highway multilayer perceptron (MLP) framework \{https://arxiv.org/abs/1605.01133}. A highway

network allows for hundreds of layers where information can flow directly between neurons of different layers. It is based on long short-term memory (LSTM) neural networks in its “gates” that control how much of the output activation and how much of the raw input to pass through the layers. In contrast to DeepBind, which involves one convolutional layer, DeepMotif has several convolutional layers. When trained on the same dataset as DeepBind, DeepMotif achieves a median AUC of 0.951 over DeepBind’s of 0.931 \{https://arxiv.org/abs/1605.01133}.

Both DeepBind and DeepMotif compute TF binding preferences based on position weight matrices (PWMs), but they do not take into account low-affinity binding sites or repeat sequence symmetries. A new model, TFImpute, takes these into account through multi-task learning (MTL) \{28234893}. This provides a more accurate model of TF-DNA binding specificity; as TFs have only been profiled under limited conditions, TFImpute could shed insight on gene expression under unique conditions. Multi-task learning models the way humans learn by exposing the model to related tasks with the goal of getting the model to perform better on one general task or metric. TFImpute uses MTL by combining combinations of cell lines and TFs into continuous input vectors. TFImpute has comparable performance to DeepBind and DeepMotif on ChIP-seq data but achieves greater accuracy on TF-cell line combinations without ChIP-seq data. However, there was greater variance in the confidence of specificity prediction on combinations without ChIP-seq \{28234893}.

Although deep neural networks can generate accurate predictions for a wide range of applications, complicated networks can sometimes be much more difficult to interpret than simpler models. Great care must be taken when selecting model parameters and preprocessing training data, as well as understanding the structure of the chosen network and how information flows through it. As computational power becomes more accessible and experimental methods improve, we will have greater amounts of high quality data to work with. Such trends will expand our need for deep neural networks to better elucidate the complicated biological networks studied in systems biology.

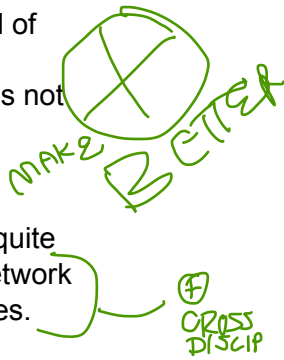
3.3 ***Approaches to understand the biological meaning of network features (WUM):***

It is common for a reader, upon encountering a visual depiction of a large network, to be struck by its incomprehensibility and struggle to make sense of the “hairball.” Fortunately, it is not necessary to comprehend the full detail of the network to interpret it biologically. Here are three fruitful alternatives, in increasing order of sophistication: focusing on one subnetwork at a time, attending to summary statistics about network properties, and contrasting a network with appropriate comparisons.

Much of the true impact of assembling full biological networks simply follows from organizing network information in a way that makes it easy for domain experts to extract the subnetwork of the graph relevant to them. The biologist studying the molecular mechanisms of some protein may care a great deal about which proteins interact with their protein of interest, as well as perhaps which proteins interact with those, and so forth. This commonplace approach to gleaning some knowledge from a network does offer some practical motivation for assembling biological networks even if it does not take full advantage of network science as a branch of mathematics.

Summary statistics about network properties are more tractable than the fully detailed network,

but are ultimately difficult to interpret without context. The distributions of node degree and of the distances between nodes, for example, offer broad, quantitative descriptions of the connectivity properties of a network and can be seen at a glance. In isolation, however, it is not clear which aspects of these distributions or other properties are biologically relevant.



The richest understanding of a biological network comes from contrasting a network with appropriate comparisons. Identifying or constructing the appropriate comparisons can be quite difficult. To understand normal biology or disease, it is often of interest to identify which network properties are under evolutionary selection or have been perturbed from their healthy states.

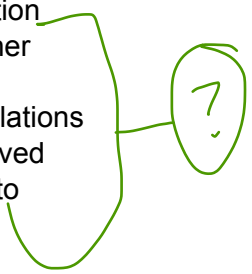
When assessing evolutionary forces acting on a network, it is important to consider how the generative processes of evolution neutral mutation affect network properties. For example, old genes beget new genes through duplication events, leading to protein products with similar binding partners as their ancestors. This neutral process leads to a characteristic “scale-free” distribution of edges within protein-protein interaction networks. Without knowledge and correction for this neutral process, an observer might incorrectly conclude that the scale-free distribution of edges within protein-protein interaction networks represents an independently evolved network property.

The identification of synergistically- or redundantly-acting mutant gene pairs in cancer is a common example of a search for selective pressures on network properties. If two genes are co-mutated more frequently than expected by chance, this suggests synergy between them. Conversely, if two genes are mutated less frequently than expected by chance, this suggests redundancy between them. A challenging task is to calibrate what sort of co-mutation frequency would be expected “by chance.” A popular approach is to generate an empirical null distribution as a series of hypothetical patient-mutant gene sets, simultaneously preserving the distribution of number of mutations across patients and genes as in the original cohort, but with permuted connections between mutant genes and patients.

3.4 Causal inference about network properties (WUM)

How does some trait affect the fitness of an organism? Which alterations contribute to disease pathology? These causal questions are among the most general and essential questions in biology. Network theory gives us, at a minimum, the vocabulary to pose these questions about network properties. More substantially, network theory offers techniques that attempt to answer these questions, some of which will be reviewed here. However, despite significant progress, causal inference on network properties remains an unsolved problem.

One indication that some network property of normal biology is not a product of evolutionary selection is if that property is a simple consequence of neutral mutational processes. A famous example pertains to the observation that the degree distribution of protein-protein interaction networks follows a scale-free, exponential distribution. Network scholars wondered whether something about scale-free distributions might favor efficient or resilient information flows through the cell and were therefore independently evolutionarily selected. However, simulations show that known neutral processes of gene duplication are sufficient to explain the observed scale-free degree distribution, and therefore, evolutionary selection need not be invoked to explain this distribution.



In general, we will not always understand neutral mutational processes in sufficient detail to model them. When there is limited knowledge of the biomechanistic processes of neutral mutation or if the biological context is farther removed from genetic processes, an alternative approach is to derive null models using more general techniques from network theory. There are two general strategies for constructing null models: forward-generative models, which build random networks from scratch, and permutation-based models, which use an observed network as a template for random networks. Both strategies hold constant chosen foundational network properties – such as the degree distribution of a network – while varying other properties of the network in a uniformly (or approximately uniformly) random way. If the network properties of the observed network significantly differ from those of the null networks, they are considered to be more likely to be fundamental to the network and therefore stronger candidates as relevant for biology and disease.

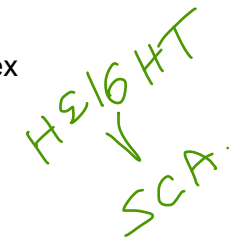
Unfortunately, there is no automatic process for selecting which network properties to hold constant when constructing null models. For example, the fact that the mammalian brain divides into two hemispheres is a foundational property of the brain that has a dramatic impact on network properties. If this inherent hemispheric structure in the brain is not taken into account, then many properties of human neural networks will incorrectly appear significantly different from null even if it were the case that they merely represented random perturbations from this hemispheric structure. This example illustrates the general principle that the fundamentality or causal impact of network properties are extremely difficult to infer and cannot be solved by any one network algorithm.

These limitations are certainly not unique to network theory, but network theory does suffer from a peculiar additional barrier in causal inference: In other areas of science, interventional experiments can definitively establish causation; whereas, it is not possible to experimentally perturb a system's network properties without perturbing its individual elements, which must always compete with the network properties as an explanation for some experimental effect.

4. APPLICATIONS

4.1 Network medicine: clinical application of molecular interaction networks (PDM):

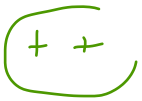
Complex diseases are conditions understood to have multiple determinants of severity, that include genetic and environmental risk-factors \{18523454}. Complex diseases include prevalent conditions like heart-disease \{22733336}, asthma \{21281866, 20860503}, autism \{21614001}, schizophrenia \{11976442,25056061}, diabetes \{27398621}, and cancer \{25109877}. Single or multiple effectors in the same molecular pathway may cause a complex disease, or a disease may result from a more distributed network effect with multiple involved pathways {24287332}.



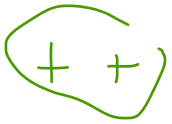
Even for so-called 'single-gene disorders' – diseases that are understood to be caused by a single mutation of a single gene -- the manifestations and severity of disease may depend a network process. For example, cystic fibrosis is a congenital lung disease caused by a defect in the CFTR membrane protein channel, but the severity of the condition may depend on an associated miRNA regulatory network \{22853952}, and on the presence of disease-modifier gene mutations \{16723978, 19242412}.

Gene set enrichment analysis (GSEA), and other forms of pathway analysis address the possibility of pathway driven diseases directly \{16199517,19033363}. Pathway analysis reveals that genetic variation in patients with autism affects many genes, but these genetic variants appear to organize into relatively few functional pathways \{24768552, 27479844}. In diabetes, many of the genes in the same pathway as the transcriptional activator PGC-1α have independently been associated with diabetes \{12808457, 27094040}. These results suggest that it may not be possible to fully understand such conditions except in the context of a network of interacting elements.

Network interactions between molecular contributors is sometimes measurable as an epistatic effect, even when the involved pathways and interactions themselves are not necessarily known \{24572353}. In cases where the source of these interactive effects between molecules is not known, subsequent identification through network-based analysis may be possible \{27708008}.



Networked based analyses have revealed shared molecular pathway alterations among diseases that were once thought distinct. Calcium-channel pathway mutations are shared by 5-different different psychiatric conditions: autism spectrum disorder, attention deficit-hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia \{23453885}. Cancers that are thought to be distinct based on organ system may share similar underlying gene and pathway alterations \{25109877}. Our understanding of relationships between diseases may be reorganized, by thinking according to network definitions of disease, rather than established disease definitions \{17502601}.



Knowledge of molecular network architecture in health and disease may also lead to disease treatment. A network approach to drug discovery allows researchers to identify new target molecules through their network interactions, and minimize side-effects by identifying the relationships between interacting molecules \{23384594}. The principle of multi-drug therapy is to address the multiple networked molecular contributors to disease -- and has led to successful management of HIV, depression, and some forms of cancer \{28697253, 22579283, 15688074, 27404187}. The bioengineering of interaction networks may be able to restore function to patients with certain diseases. An engineered gene-network restored thyroid function in a mouse model of Grave's disease \{26787873}.

4.2 Network techniques in cancer genomics (JZ):

Molecular networks have particular relevance to cancer biology. Using a pathway or network based approach to analyzing mutational patterns, cancer types may redefined or subcategorized. This approach, when performed as part of a broad molecular profiling strategy, has defined novel cancer subtypes for many cancers including breast cancer \{23000897}, melanoma \{26824661}, lung cancer \{25079552}, and kidney cancer \{26536169}. Significantly, the only route to diagnosis of metastatic cancer of unknown primary origin may be through analysis of the patterns of activity and cross-talk defined through molecular profiling \{25140961}.

Regulatory networks may provide deep functional annotations to more accurately evaluate mutation impact and prioritize key mutations in cancer. For example, network centrality information has been used by researchers to pinpoint key cancer mutations \cite{netsnp and funseq2}. Transcription factor (TF) and RNA binding protein networks may also provide insight

to explain disease-specific expression patterns and help highlight key cancer regulators. For instance, by combining large-scale expression profiles from cancer patients with TF networks identified by ChIP-seq experiment, it is possible to identify important TFs that drive tumor-to-normal differential expression \{rabbit 26056275, 28000771}.

Integration of diverse sources of biological network data may be used to reveal novel cancer biology [Figure 7]. Integration of TF-gene, miRNA-gene, and protein-protein interaction network data has been used obtain a systems-level view of various diseases, including cancer. This integration of diverse network information may be used to highlight key genes and mutations associated with tumorigenesis. Unlike mutational frequency based methods, which require sequencing data from large cohorts to achieve satisfactory statistical power, scientists are able to obtain a global view of mutational effect for multiple genes through network propagation techniques \cite{DriverNet 23383675, VarWalker 24516372, HotNet2 25501392, NBS 24037242, and TieDIE 23792563}. Such methods have been successfully applied on moderately sized patient cohorts to identify cancer related genes.

Network associations may yield new cancer therapies. For example their is great interest in the molecule CMTM6 because it has been shown to interact with the molecule PD-L1 and regulate PD-L1 expression. PD-L1 itself helps regulate the body's immune response to cancer cell surface markers, and \{28813417}. Thus, perhaps CMTM6 will prove similarly useful as a regulatory target. Knowledge of these pathways may result in development of new cancer therapies, and combination drug therapies that reduce the risk of developed resistance to cancer therapy \{27433843,25838373}.

4.3 Cross-disciplinary comparisons provide insight into molecular interaction networks. (PDM):

We may learn more about the mechanisms and function of molecular networks through cross-disciplinary comparison to networks found in other natural and engineered systems [Figure 8].

A comparison of the transcriptional interaction network of the bacteria *Escheriae coli* to the call graph of the Linux operating system demonstrated that the transcriptional network in *Escheriae coli* has a robust architecture, with many network elements sharing overlapping function \{20439753} [Figure 8.a.]. Conversely, the Linux call-graph is built on frequent reuse of many basic operating functions. An analysis of biological protein-DNA and protein-protein interactions in both *Saccharomyces cerevisiae* and *Escheriae coli* to internet connectivity networks also favored the robustness of the biological networks \{24789562}.

Rieckmann et al. recently conceptualized the human immune system as a social network. By mapping a social network architecture based on cytokine 'messages' between cells, these researchers demonstrated unexpectedly close relationships between immune cell types \{28263321}. For example, neutrophils and naive-B-cells were unexpectedly closely related, as were natural killer cells and memory T-cells \{28345632}. It's intriguing to think that the discovered proximity of relationships in this 'small-world' network may reflect how immune cells interact within the compartments of the human body \{28418389}.

Metabolic networks have been described as a type of ‘scale-free’ network, meaning that the network is self-similar at each scale, with the degree of nodes following a power law. Metabolism appears organized around two central hubs -- pyruvate and acetyl-CoA \{15729348}. This is similar to how already well-used airports are likely to gain additional flight routes due to the efficiencies in airline travel that are gained by travelling through a network hub \{10.1038/nphys489, 15911778}.

Like metabolic networks, protein-protein interaction networks are also often thought of as ‘scale-free’ networks, following this same rich-get-richer principle \{*doi:10.1038/nphys209*}. However, researchers have also suggested that protein-protein interaction networks may be more similar to geometric networks based on their network topology \{15284103}. Electrical grids are connected based upon the existing geographies of cities, and wireless mesh networking similar connects electronic devices based on spatial proximity. The observation that protein-protein interaction networks appear to have geometric network topology, may be related to the spatial organization molecules within the cell, as determinant of their interactions \{15284103, 25985280}. Geometric constraints within cells may also provide bio-inspired templates for efficient generation of geometric graphs. Such a possibility was demonstrated through comparison of the growth of the single-celled organism *Physarum plasmodium* to the rail system in Tokyo \{20093467}.

5. CONCLUSION:

We began with an overview and introduction to how molecular interaction networks have played an irreplaceable role in the development diverse scientific fields including molecular biology, medicine, and network science. We surveyed how network topology and network dynamics may be used to derive insight into human biology and human disease. The time-dependency and computational capacity of interaction networks offer a means of maintaining homeostasis, but these same networks may also serve as the sensor and driver of common diseases.

We discussed two promising algorithmic approaches to identifying novel molecular associations: network propagation algorithms that seek to identify important associations between molecules through a diffusion-type process, and machine learning techniques, including deep learning models, that may identify novel network structure through sophisticated pattern recognition performed on markers of molecular interaction. Related to this discussion of network algorithms, we provided some viewpoints on how the study of interaction networks can benefit from network comparisons. These comparisons can be made via a null model of interaction -- a random generative process -- or in comparison to other biological or nonbiological networks.

Our discussion of molecular interaction networks concluded on the topic of applied uses for molecular interaction networks. Applications in medicine have resulted greater knowledge of disease, and new disease treatments. We highlighted the case of networks in cancer, as a particular set of diseases that have benefited from applied network science. The use of molecular interaction networks to make cross-disciplinary comparisons, has lead to greater understanding of networks in wide-ranging fields of study.

We hope to have given the reader of sense of the strategic significance of network analysis techniques and interaction networks. These authors hold the strong conviction that because molecular interaction networks are the lowest common denominator in many higher-order

biological systems, network analysis techniques will be a critical component of future advances in molecular biology and medicine. *These authors further believe that there will be cross-disciplinary advantages to investigation of molecular interaction networks, propelled by the need for the adoption of new network techniques to analyze large data sets, and by the need to integrate diverse sources of information.*

6. SUMMARY POINTS:

1. Molecular interaction networks represent the base-layer of function for many higher-order biological systems, and have contributed to the development of knowledge in biology, medicine, and data science.

2. Abstract network representations provide a useful platform to model network behavior, however, not all interactions can be inferred without molecular structural information.

3. Molecular networks are not static, but evolve over time and space, and this evolution enables function in both health and disease.

4. New algorithms for understanding molecular interaction have revealed novel molecular relationships. Promising techniques include network propagation, and neural network based deep-learning models.

5. The significance of a molecular interaction network requires a comparison standard -- a null model (either physiologic or randomly generated), or a cross-disciplinary comparison can serve as such a comparison standard.

6. Many disease processes arise through pathway or network phenomena, and require an understanding of network properties to understand their pathology and identify treatment strategies.

7. Cross-disciplinary network comparisons contribute insight into molecular network structure and function.

7. FUTURE PROSPECTS:

1. Challenge of identifying appropriate null comparisons for molecular interaction networks. Possible null comparisons include random network rewiring, random generative processes, and cross-disciplinary network analogies.

2. Incorporation 3 dimensional structure, and time dependency (network logic, network rewiring) into network models.

3. New application of network algorithms to refine network predictions, including machine learning techniques, and network propagation algorithms.

4. Designing efficient, scalable algorithms for large search spaces that provide accurate approximations to actual network behavior.

5. Defining scalable approaches for integrating diverse molecular data sets, including functional genomics data.

6. Translational research, applying techniques to medicine, and scaling solutions for clinical data, including correlation with clinical phenotypes.

7. Increased experimentation with network engineering and network intervention as a means of disease treatment.

8. Expansion of cross-disciplinary network science efforts, for example molecular epidemiology (intersection of social networks, molecular networks, and epidemiology), molecular phenotypic pathology (intersection of pathology and molecular networks).

9. Redefinition of disease by molecular phenotype and molecular pathology will require substantial pathway and network analysis.

10. Identifying appropriate validations for the predictions of network analyses on a genomic scale.

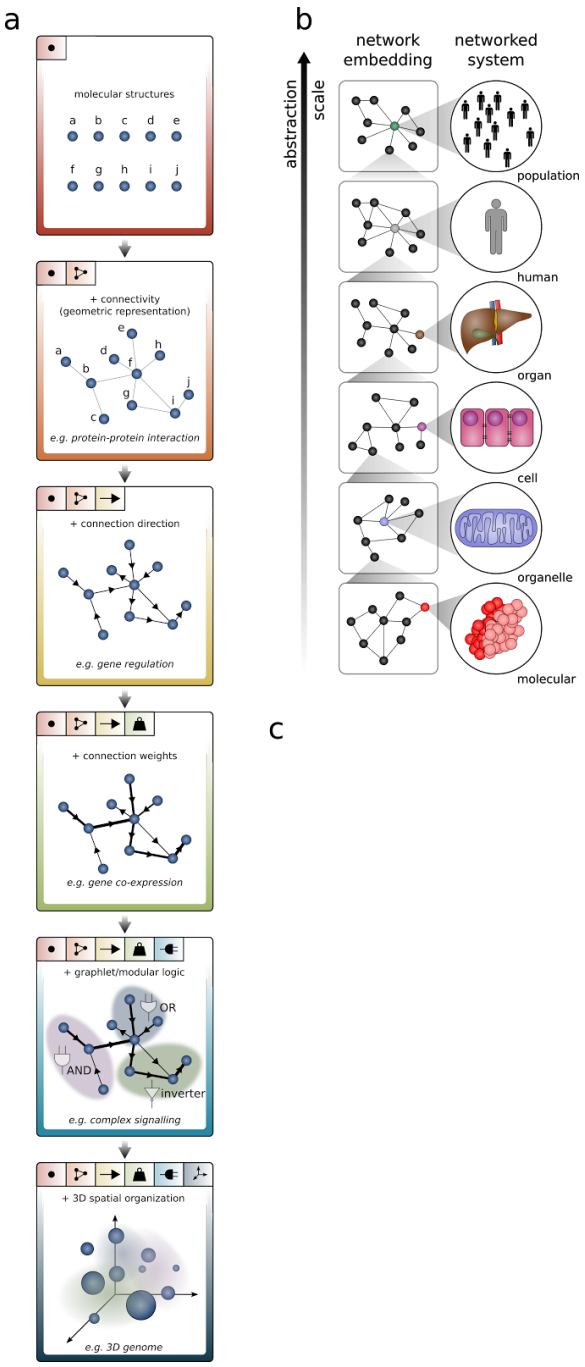


Figure 1: a) b) c)

Figure 2:

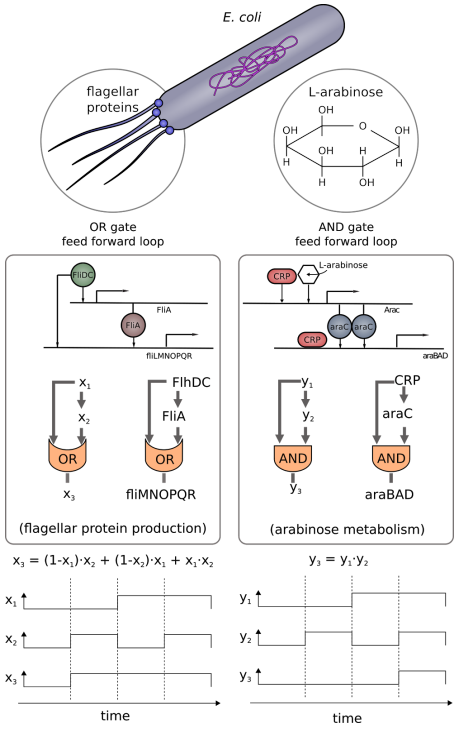

E. coli
flagellar proteins
L-arabinose
OR gate feed forward loop
AND gate feed forward loop
FlhDC
FliA
fliLMNOPQR
CRP
L-arabinose
Arac
araC
araBAD
FlhDC
FliA
OR
OR
fliMNOPQR
(flagellar protein production)
CRP
araC
AND
AND
araBAD
(arabinose metabolism)
$x_3 = (1-x_1)\cdot x_2 + (1-x_2)\cdot x_1 + x_1\cdot x_2$
$y_3 = y_1\cdot y_2$
time
time

Figure 2: a) b) c)