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 resilient, also make some diseases difficult to treat.

Network theory is a well-developed branch of mathematics that organizes and analyzes 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. Network analysis techniques first developed to examine transportation networks, or social networks, or communication infrastructure, provide insight through application to biomedical data.

The application of diverse network analysis techniques to molecular data reveals pathways of cellular function distributed over molecular networks. Key molecular network elements and architecture may lead to disease when disrupted. The comparison of network interactions in health and disease provides a platform for disease diagnosis and design of medical therapies.

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):

A great diversity of networks are relevant to the field of biomedicine. Social networks model human interaction and may help explain pathways of disease transmission. Layers of neurons in the brain process sensory information, and the layered architecture of neuronal networks inspired the artificial neural networks that are used to identify patterns in biomedical data \{28728020}. The circulatory system is a branching network of vessels that connects organs in the body. Vast networks of interacting molecules, in particular, are foundational to human health and disease. Molecular networks form a functional base layer for a number of higher-order biological networks [Figure 1.b]. Transfer of genetic information, cellular communication, and human metabolism are all mediated by complex pathways and networks of molecules.

Networks are a powerful framework for understanding molecular interactions because of the breadth of network analysis techniques developed across diverse disciplines. Novel network analysis techniques like HotNet \{25501392,21385051} use algorithms first developed in the context of understanding belief propagation in social networks \{Reverend Bayes… Judea Pearl 1982} to annotate function in molecular networks. Machine learning techniques like the deep neural network DeepBind, apply techniques refined for use in computer vision \{ImageNet, Hinton} to generate accurate network topology predictions from genomic sequences \{26213851}. Cross-disciplinary comparisons between networks reveal that the gene-regulatory network of *E. Coli* is built for robust function in comparison with a computer software network that prioritizes economy and reuse \{cite}. In comparison to a social network, apparently distant connections between immune cells may in reality be closer than appearance through mutual aquaintance \{28263321} and 'cross-talk' between immune cells may modulate the body's immune response \{24923297}.

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 large-scale molecular 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 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.

Network analysis of biomedical data is not just a research technique, but has also contributed to advances in understanding and practice in 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.

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 for 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.

1.3 Making sense of complexity in biomolecular networks (WUM):

Complex biomolecular networks may be incomprehensible in their raw, complete form. Fortunately, there are at least three ways to make sense of them. Most straightforward, networks become comprehensible by focusing on only some portion of the full network. A more scalable approach is to compute summary statistics about the network. Alternatively, networks can sometimes best be appreciated through comparison to other networks.

Networks are like maps in that both organize local information in global context. This is analogous to a map of the world, where the architecture of cities cannot be appreciated at the scale of countries or continents. Large, complex biomolecular networks are best visualized with either reduced detail, restricted scale, or both, except when demonstrating the size of a dataset. For example, although metabolism is an extremely complex process (See Figure WM.A below), glycolysis – the core sub-graph of metabolism (Figure WM.B) – is simple enough for a dedicated high school student to appreciate in an afternoon while rich enough to convey principles of metabolism. In Section MTG.X we use logic gates as a case study to illustrate the interpretative utility of subnetworks. Premier online databases of biomolecular networks, such as the KEGG Pathway database \{27899662}, support interactive visualizations of networks.

A second way to understand a biomolecular network is through its summary network properties. Stanley Milgram famously discovered that between any two residents of the United States he studied, there are on average six degrees of separation \{The Small-World Problem, Milgram}. The short average path length of the American social network is an interesting property that helps us to appreciate how people are connected to each other, and how ideas and infections can quickly spread. In the human protein-protein interaction network, one study found that the average path length is around 4.85 \{16169070}. This connectivity between proteins helps us appreciate why so many different proteins may be relevant to a given human trait or disease.

Some of the most interesting insights about biomolecular networks come from making comparisons between networks. Comparing a biomolecular network against randomized networks helps us appreciate which properties are fundamental to a network, vs merely expected by chance (See Section WM.X). Contrasting a biomolecular network between healthy and diseased samples highlights changes that may be relevant in disease pathogenesis (See Section DL.X for an application to cancer). Comparing biological networks with man-made networks that have been designed for some function can inspire us to wonder, with due caution, whether the biological network has been evolutionarily designed to perform that function (see Section PDM.Y for examples but WM.X for challenges in making inferences about evolutionary forces in networks).

These three approaches to understanding networks also inspire ways we might improve our methodology to make networks more intelligible still. Biomolecular network scientists may improve visualization by borrowing techniques from interactive cartography. Combining the second and third approaches: we can better understand the meaning of biomolecular network summary statistics by finding and making more apt and more diverse comparisons between networks. It is understandable that biomolecular networks, given their enormity, have something of a reputation for inscrutability among outsiders. What else is understandable?

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 [Figure 3.a.]. 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.b.]. 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 are decades old. Prior efforts to understand network dynamics compared transcription factor-gene networks in varying conditions \{15372033}. However, the scope of these efforts were limited by data-availability. 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 a variety of cellular contexts. It remains difficult to measure the dynamic nature of biological networks. However, advanced biomolecular assays can provide clearer insight into how genes and proteins operate in point-in-time snapshot. Researchers may then stitch these snapshots together to answer complex, time-dependent 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 particular, it has been shown that regulatory changes in transcription factor-target networks may account for species differentiation \{17690298, 20378774, 21253555, 23198090}. However, researchers have also attempted to interpret network rewiring at much shorter timescales. It is possible to introduce an artificial perturbation into a network and examine the rewiring that results. One study of a bacterial gene network showed that a single perturbation can affect gene expression by four orders of magnitude to the scale of perturbation, and alter up to approximately 70% of the transcriptome \{26670742}.

Rewiring is often the result of genetic mutation. A single mutation placed at a regulatory protein binding site can alter binding specificity, perturb its interacting neighbors, and consequently, have a detrimental downstream effect on the whole network. Naturally, many researchers have attempted to measure rewiring to infer the consequence to disease phenotype.

For example, cancer mutations can affect both downstream and upstream rewiring of the gene-regulatory network, altering cell-signaling and non-cancer gene expression \{26388441,26388442}. Measuring rewiring (i.e., target changing) of a transcription factor-gene network involves comparison of networks in two states: the reference (healthy) state and the evolved (diseased) state networks. Measuring the extent to which a gene perturbed in a network has revealed tumor drivers and genes associated with patient prognosis \{27145341}. Regulatory interconnection between genes can be represented as the gain, loss, or retention of molecular interaction. As a result, network rewiring can change gene hierarchy, promoting or demoting 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 design more complex models to test the effects of cancer mutations. It could prove to be an excellent tool for both performing a high-throughput screening of network perturbation and experimentally validate the results of rewiring obtained via an integrative approach.

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

Rewiring may be viewed as an irreversible temporal evolution of biomolecular network. However, when viewed at a much shorter time-scale, biomolecular dynamics can be understood as concerted and responsive changes in a biomolecular network. Pairs of regulatory molecules can work collaboratively, competitively, or redundantly. More complex function -- like the integration of time varying hormonal signal, or a conditional cellular response to an environmental change -- is enabled through the dynamic behavior of molecular networks. Molecular networks may even compared to logic gates \{25884877}, with spatiotemporal information revealing their mode of operation.

At the evolutionary time scale, biological networks such as protein-protein interaction networks, have evolved to maximize network efficiency, functionality, and stability. 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, allow the use of strong promoters which shorten the response time of stimuli-induced gene expression regulation. The auto-repressive nature of these motifs allows cells to quickly attain stable protein-product concentrations, and reduce cell-cell variation in proteins levels \{12417193}.

Another frequently observed motif in gene regulatory networks is the feedforward loop. Unlike direct stimuli that generate a rapid response, feedforward loops with AND gate logic require more persistent stimulation to activate both input components, thus filtering out brief spurious pulses of signal. Combinations of network motifs enable more precise control of biological systems, including the temporal order of gene expression and circadian oscillations. \{17510665}.

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 Network prediction using machine learning and neural networks **(Holly Zhou):**

Network prediction methods have evolved in parallel with the evolution of large-scale biological experimentation. Experimental molecular interaction data contains both false-positive interactions and false-negative interactions \{STRING?}. Predictive algorithms attempt to identify these false-positive and false-negative cases, and so address the limitations of experimental methods. For the well-studied case of protein-protein interaction data, diverse predictions methods include predictions based on gene ordering and genetic sequence \{9787636}, network topology \{18451861}, Bayesian inference and machine learning methods \{14564010}, measurements of structural similarity \{23023127}, and text mining \{STRING}. Network prediction methods can be combined in an attempt to yield more accurate predictions, and a large body of literature is devoted to improving network predictions \{24396273}.

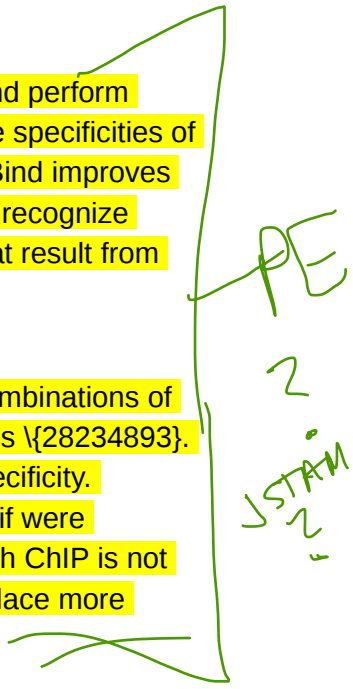
Machine learning methods, and neural networks in particular, have become popular methods for network prediction. Machine learning methods can predict relationships in networks without necessarily requiring strong assumptions about underlying interaction mechanisms \{27474269}. Dimensionality reductions makes large genomic data sets more computationally tractable, and machine learning methods also allow diverse data types and a wide-variety of molecular features to be integrated to form predictions *\{26252139}.* These attributes allow these methods to scale with increasing volumes of high-throughput molecular data, and to accommodate new forms of data as they become available.

An example application for network prediction, is the identification of DNA and RNA targets of regulatory proteins. An accurate understanding of gene-regulatory networks is important for modeling networked biological processes, and for determining the impact of genomic variants -- particularly those variants in non-coding regions that do not directly affect protein structure. Predictive methods can integrate protein-sequence interaction data from a variety of sources, including protein-binding microarray and chromatin immunoprecipitation (ChIP), while also tolerating bias and error latent in these data sources.

Recently developed neural network algorithms designed for DNA- and RNA-protein interaction prediction include DeepBind \{26213851}, DeepMotif \{https://arxiv.org/abs/1605.01133}, and TFImpute \{28234893}, a deep-learning based imputation method for transcription factor (TF) binding prediction. These convolutional neural networks aim to provide better understanding of regulatory network structure, and provide tools for researchers to prioritize mutations by their impact on protein binding sites \{27197224}.

DeepBind and Deep Motif take sequencing data from high-throughput experiments, and perform convolution of sequence-based protein binding motifs, in order to predict the sequence specificities of DNA- and RNA-binding proteins \{26213851, https://arxiv.org/abs/1605.01133}. DeepBind improves upon prior motif-scanning algorithms by taking into account RNA-binding proteins that recognize secondary or tertiary structural elements. It also recognizes higher-order structures that result from competitive or synergistic effects of protein binding \{26213851}.

In order to predict transcription-factor binding sites, TFImpute takes input data from combinations of cell lines, and also considers low-affinity binding sites and repeat sequence symmetries \{28234893}. These features are designed to provide a more accurate model of TF-DNA binding specificity. Improvements at TF binding site prediction for TFImpute over DeepBind and DeepMotif were particularly notable in sequencing from cell-types for which protein binding data through ChIP is not available. This suggests an application for predictive computational approaches to replace more expensive experiments that may have limited availability.



As experimental methods improve and evolve, computational biologists can expect to have greater quantities of high quality data to work with. Predictive algorithms that are able to integrate diverse data sources, while also scaling with increasing dataset size, are those that will be most helpful in elucidating the complicated biological networks studied in systems biology.

***3.3 Causal inference about network properties (WUM)*:**

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

When assessing selective forces acting on a network, it is important to consider how the generative processes of evolution 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.
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 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):

Some diseases, like sickle-cell anemia, are thought to be caused by single mutations or alteration of a single genetic locus \{28423290}. Complex diseases are conditions understood to have multiple determinants of severity, that include genetic and environmental risk-factors \{18523454}. This is

similar to how complex traits like height are thought to arise from the interaction of multiple genetic loci \{25282103}. 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 those 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 known \{24572353}. Epistatic interaction is contribution to a phenotype from interplay between multiple molecular partners \{15266344}. Epistatic effects are an important reason why molecular changes cannot always be studied in isolation from their network interactions: interacting molecules may modulate the relative impact of their binding partners. The effects of epistatic interaction on disease phenotype, highlights a need for network analysis to understand disease pathogenesis -- in cases where the source of these interactive effects between molecules is not known, subsequent identification through a systems-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}. This overlap of molecular phenotype among diseases that were once thought distinct, may change how we think of disease and diagnosis. Rather than relying on established disease definitions, our understanding of disease may be shaped by a network definition of disease. Relationships between diseases may be better understood in the context of a global 'diseaseome' \{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. 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. The integration of data from multiple molecular networks, allows cancer mutation to be understood from the perspective of aggregated effect on multiple networked subsystems [Fig 7]. Unlike mutational frequency based methods, which require sequencing data from large cohorts to achieve satisfactory statistical power, scientists can obtain a global view of mutational effect through multiple networked genes. Methods that include network propagation techniques can help identify these aggregated mutational effects \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]. The comparison of a designed system with an evolved system may reveal the evolutionary pressures that shape complex biology. Network attributes that vary among among biological and engineered networks may highlight functional network architectures. Through such comparisons, advantages to evolved biological systems may be identified that can improve human-engineered systems through biomimicry.

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}. Just as flight options are most easily expanded by connecting to an already well-connected airport, pyruvate and acetyl-CoA may function as hub-metabolites, facilitating molecular transitions between biochemical pathways.

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 a 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, and 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 amplify important associations between molecules through a diffusion-type process, and network prediction 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 biological data.

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:

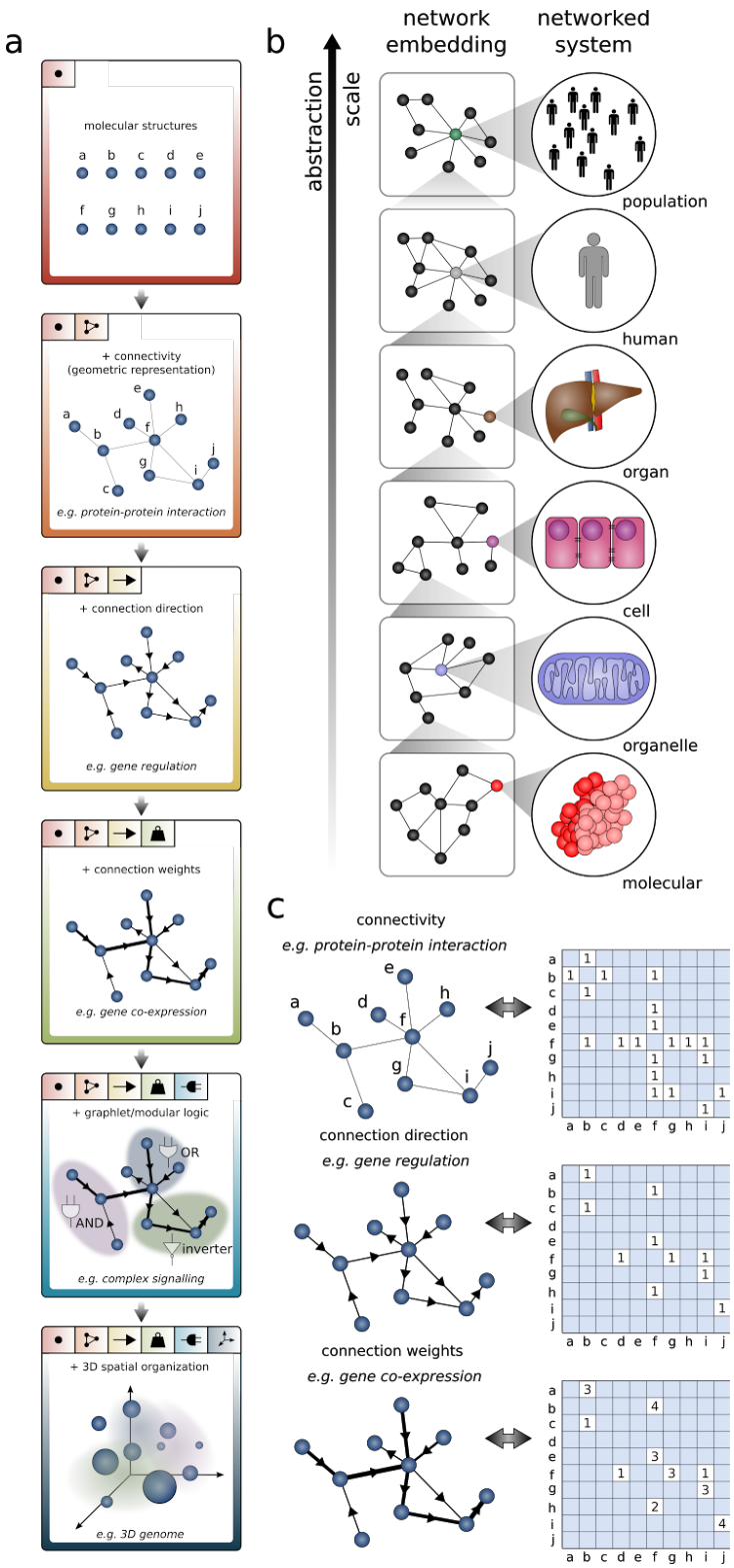

a
molecular structures
a b c d e
f g h i j
+ connectivity
(geometric representation)
e.g. protein-protein interaction
+ connection direction
e.g. gene regulation
+ connection weights
e.g. gene co-expression
+ graphlet/modular logic
OR
AND
inverter
e.g. complex signalling
+ 3D spatial organization
e.g. 3D genome
b
network embedding
networked system
abstraction
scale
population
human
organ
cell
organelle
molecular
c
connectivity
e.g. protein-protein interaction
connection direction
e.g. gene regulation
connection weights
e.g. gene co-expression

Figure 2:

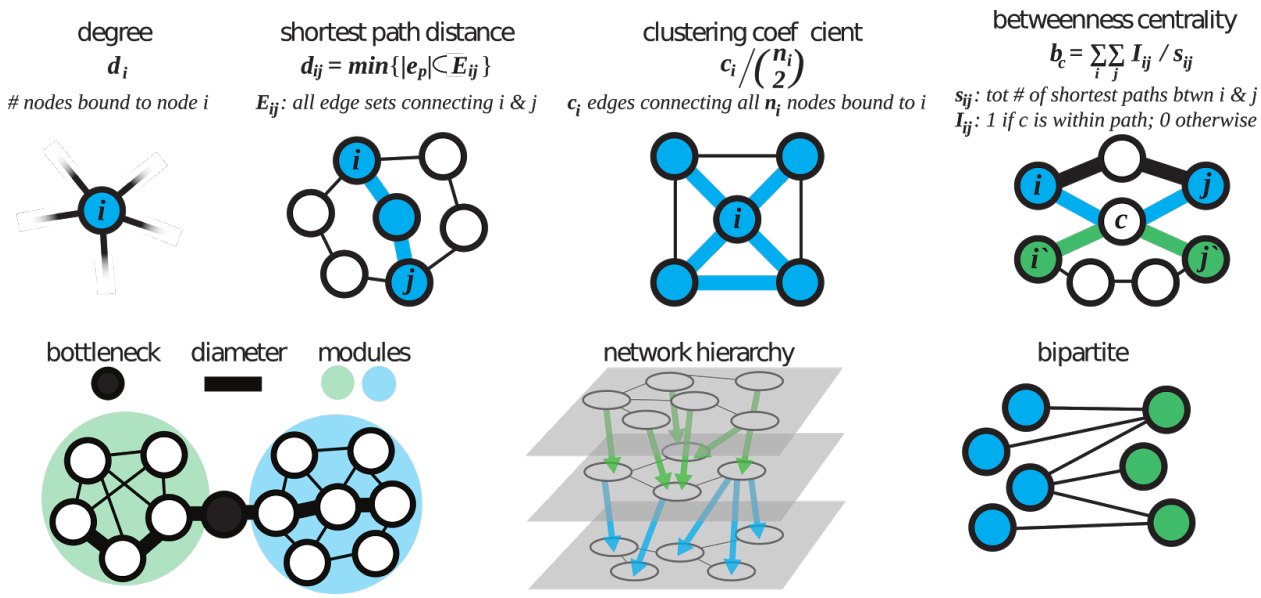
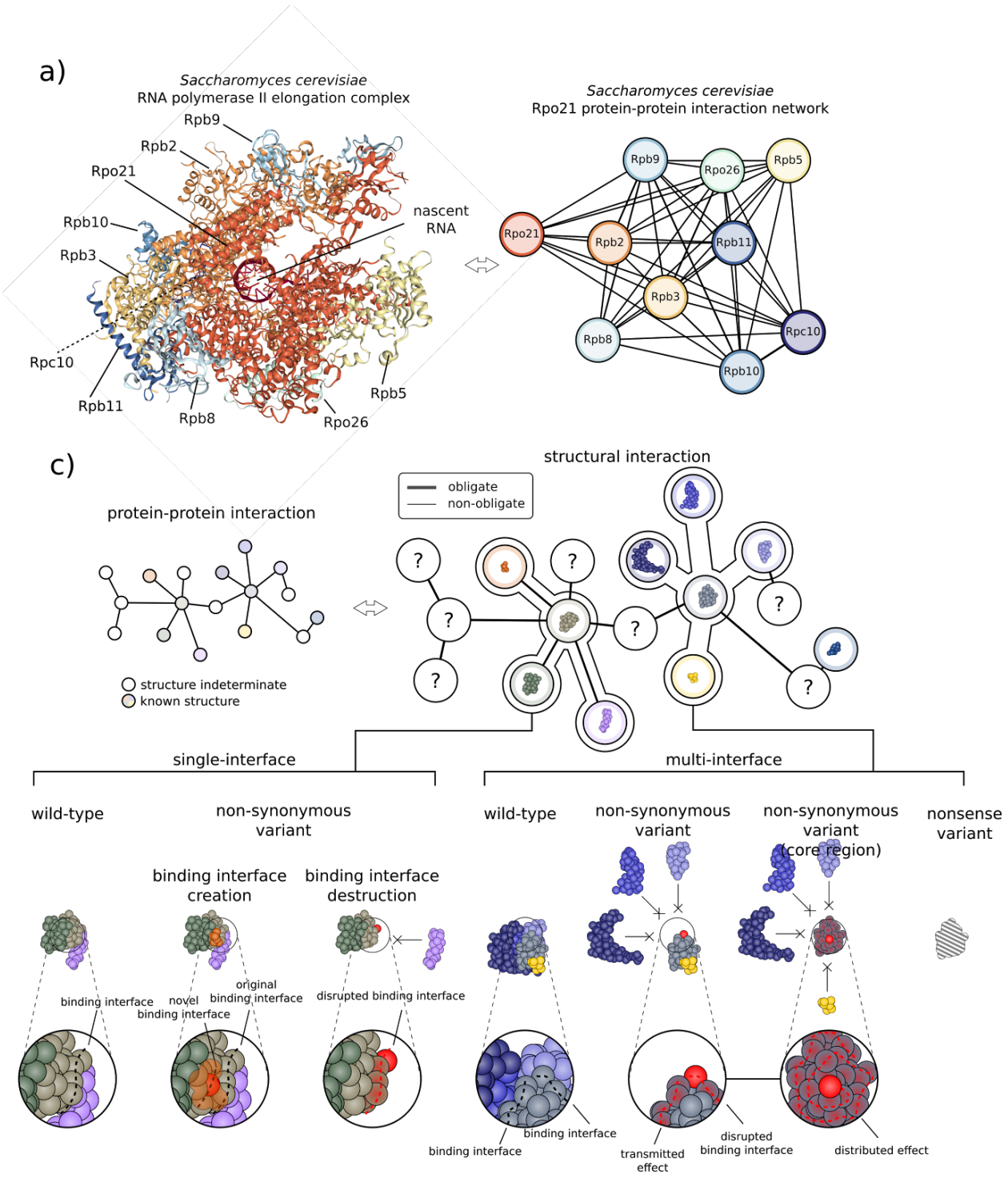

degree
d_i
nodes bound to node i
shortest path distance
$d_{ij} = min\{|e_p| \subset \bar{E}_{ij}\}$
E_{ij}: all edge sets connecting i & j
clustering coef cient
$c_i / \binom{n_i}{2}$
c_i edges connecting all n_i nodes bound to i
betweenness centrality
$b_c = \sum_i \sum_j I_{ij} / s_{ij}$
s_{ij}: tot # of shortest paths btwn i & j
I_{ij}: 1 if c is within path; 0 otherwise
bottleneck
diameter
modules
network hierarchy
bipartite

Figure 3:



RNA Pol ii S. cerevisiae:

https://www.rcsb.org/pdb/ngl/ngl.do?pdbid=1I6H&bionumber=1

http://www.rcsb.org/pdb/ngl/ngl.do?pdbid=5SVA&bionumber=1

http://www.uniprot.org/uniprot/P04050

https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=5sva

https://string-db.org/cgi/network.pl?taskId=jxinxvjwVmR5

http://www.cell.com/cms/attachment/2002982441/2011316630/gr1.jpg

https://www.google.com/search?q=RNA+pol+ii+cerevisiae+interaction+network&client=ubuntu&hs=cQ5&channel=fs&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiG9u6Dw7bWAhUl0YMKHe8GCuEQ_AUICigB&biw=1093&bih=597#imgrc=PZCwu0q0JW7j8M:

Figure 4:

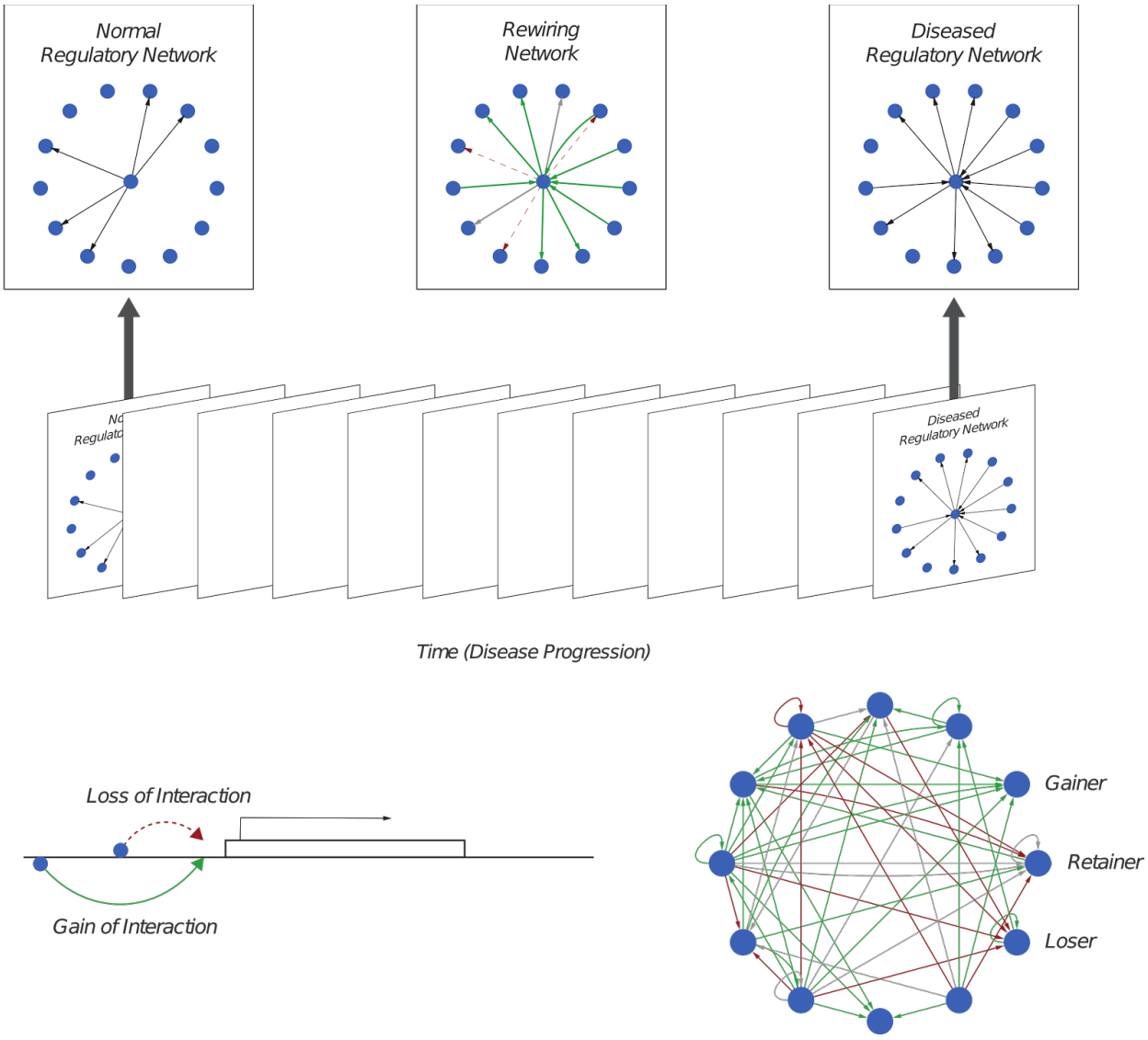


Figure 5

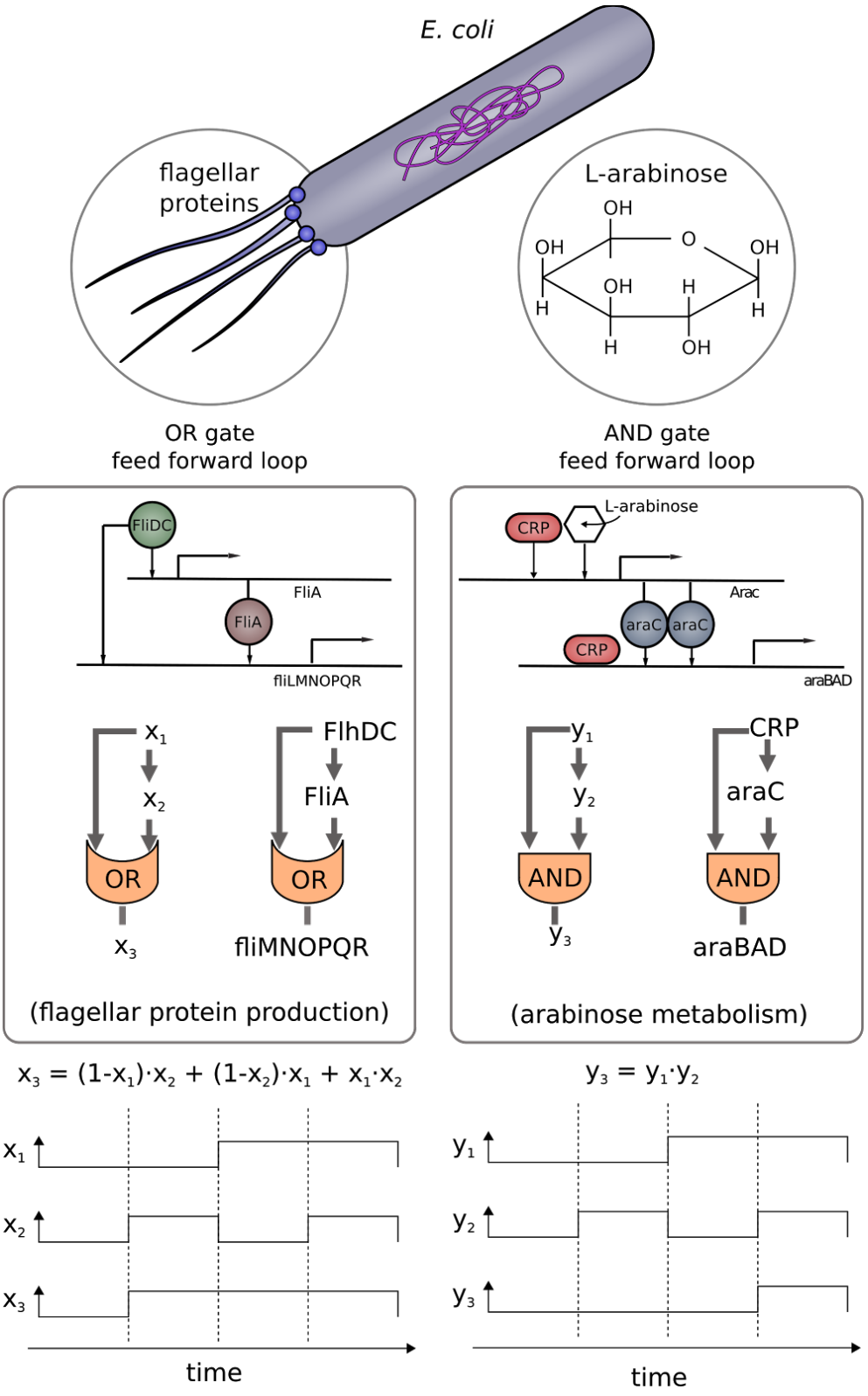

E. coli
flagellar proteins
L-arabinose
OR gate feed forward loop
AND gate feed forward loop
FliDC
FliA
FliA
fliLMNOPQR
CRP
L-arabinose
Arac
araC
araC
CRP
araBAD
x1
x2
OR
x3
FlhDC
FliA
OR
fliMNOPQR
(flagellar protein production)
y1
y2
AND
y3
CRP
araC
AND
araBAD
(arabinose metabolism)
x3 = (1-x1)·x2 + (1-x2)·x1 + x1·x2
y3 = y1·y2
x1
x2
x3
time
y1
y2
y3
time

Figure 6

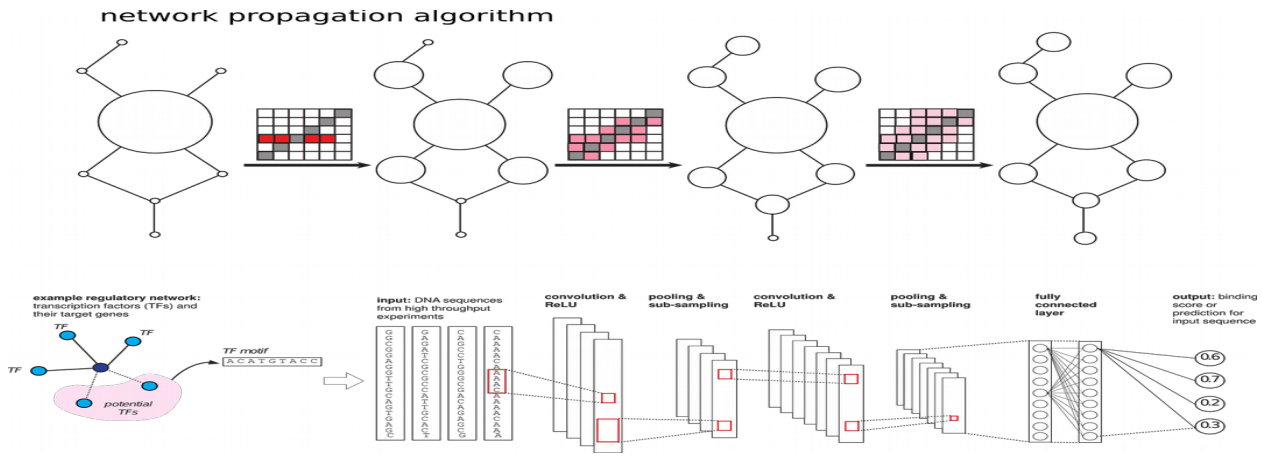


Figure 7:

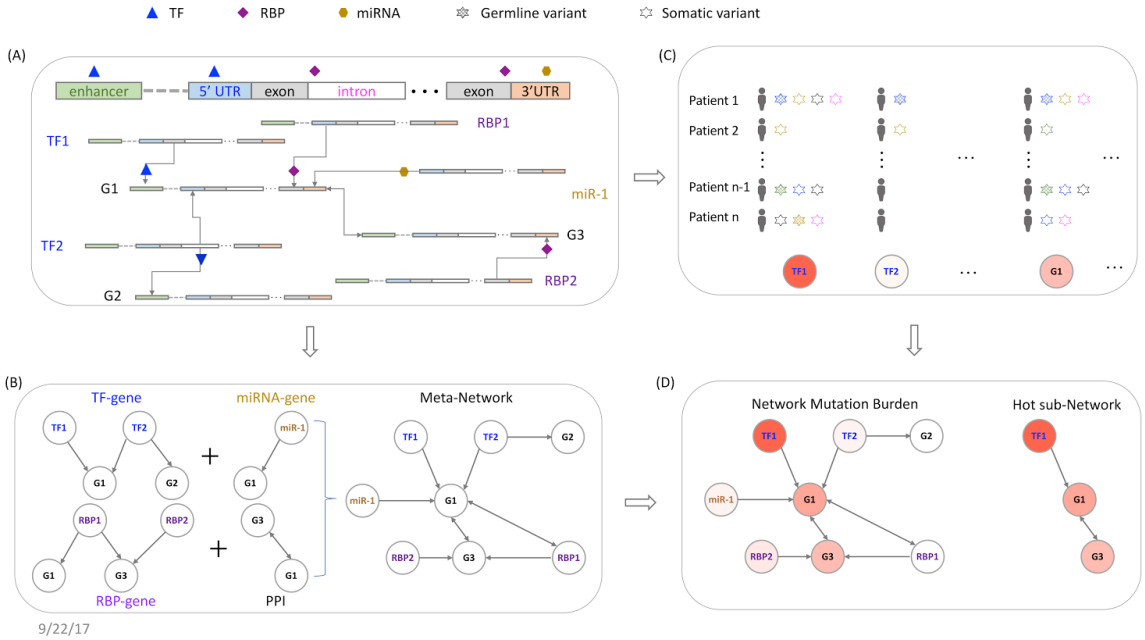
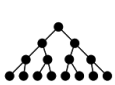


Figure 8:



hierarchical

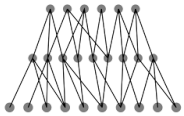
small world

scale-free

geometric random

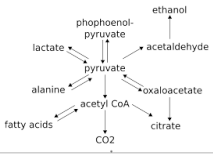
biological

transcription factor regulation

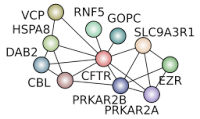


immune regulation

metabolic network

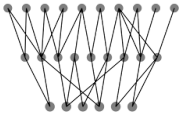


protein-protein interaction



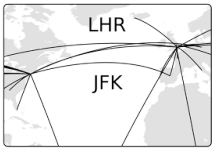
comparison

linux call graph



social interaction

airline network



electrical distribution

comparative insight

"robustness (biological) vs. efficiency and reuse (software)"

Yan et al. 2010

"Well connected despite low relatedness"

"Unexpectedly close relationships between cell types."

Rieckmann et al. 2017, Bird 2017

"rich get richer"
"oligarchy of hubs"
"rich club"

Guimera et al. 2005, Guimera et al. 2005

"proteins function in 3-dimensional space and time"

"designed and optimized communication networks"

HIgham et al. 2008