Network-based approaches for analysis of complex biological systems

Deborah Chasman\textsuperscript{a}, Alireza Fotuhi Siahpirani\textsuperscript{b,a,c}, Sushmita Roy\textsuperscript{a,c,h,*}

\textsuperscript{a}Wisconsin Institute for Discovery
University of Wisconsin–Madison, Madison, WI 53715
\textsuperscript{b}Department of Computer Sciences
University of Wisconsin–Madison, Madison, WI 53706
\textsuperscript{c}Department of Biostatistics and Medical Informatics
University of Wisconsin–Madison, Madison, WI 53792

Abstract

Cells function and respond to changes in their environment by the coordinated activity of their molecular components, including mRNAs, proteins and metabolites. At the heart of proper cellular function are molecular networks connecting these components to process extra-cellular environmental signals and drive dynamic, context-specific cellular responses. Network-based computational approaches aim to systematically integrate measurements from high-throughput experiments to gain a global understanding of cellular function under changing environmental conditions. We provide an overview of recent methodological developments toward solving two major computational problems within this field in the past two years (2013–2015): network reconstruction and network-based interpretation. Looking forward, we envision development of methods that can predict phenotypes with high accuracy as well as provide biologically plausible mechanistic hypotheses.

Keywords: Computational systems biology, context-specific network, regulatory network, network-based prediction, gene prioritization, interpretation

*Corresponding author
Email address: sroy@biostat.wisc.edu (Sushmita Roy)
Introduction

Normal and disease cellular states are the integrated outputs of networks acting at multiple levels of regulation, including the pre-transcriptional, transcriptional and post-translational levels [1, 2, 3, 4]. Advances in omic techniques are providing an unprecedented capacity to measure RNA (coding and non-coding), protein, and post-translational modification levels under different biological contexts such as time, cell states, tissues, or organisms [5, 6, 7, 8]. Systematic integration of such datasets is essential to identify molecular networks controlling normal and disease states, and, ultimately, predict complex phenotypes from molecular markers [9, 10]. This review focuses on recent efforts (2013–2015, Supplementary Table 1) in the field of network biology to address two major challenges that have emerged in the era of high-throughput biology (Fig. 1): Network reconstruction, which entails inference of shared and context-specific network connectivity from multiple omic measurements of molecular entities, and Network-based interpretation of experimental observations to guide further study.

Network reconstruction

A long-standing problem in systems biology is to infer causal, regulatory connections among genes, proteins and metabolites and to understand how connections change over time, cell types and other conditions. A network provides a natural representation of a complex cellular system with nodes representing the molecular components and edges representing different types of connectivities (Box 1A). The goal of network reconstruction (also referred to as network inference) methods is to infer such connections from genome-scale measurements of genes, proteins and metabolites from multiple conditions (Fig 1A). In this review, we focus on transcriptional networks and briefly discuss extensions to post-transcriptional, signaling and metabolic networks. We first discuss recent work in the widely studied problem of reverse engineering a single genome-wide network from large collections of gene expression data, followed by approaches to infer integrative and context-specific networks, problems of emerging interest.

Regulatory network reconstruction for one condition

Since the advent of microarrays and more recently RNA-seq to measure genome-wide transcriptomes, expression-based network inference methods have been popular for inferring regulatory networks among regulatory proteins such as transcription factors (TFs) and signaling proteins and target genes. A regulatory network model has two components (Box 1B): (a) the structure, which specifies the regulators of a target gene, and (b) the regulatory function, encoded as a mathematical function, which describes how individual and combined regulatory inputs specify a target gene’s expression. Several different mathematical functions have been proposed to relate regulatory inputs to expression out-
put including Boolean functions, ordinary differential equations, thermodynamic models, and probabilistic functions (reviewed by Kim et al. [11]).

Expression-based network inference has been approached in two ways: by reconstructing regulatory programs for (i) individual genes at a time ("per-gene" methods) and (ii) modules of genes ("per-module") (reviewed by DeSmet & Marchal [12]). While per-gene methods provide fine-grain gene-level regulatory information, per-module methods have the benefit of being easier to interpret and capture modular organization of networks (Box 1D). Some recent work combines the two paradigms [13], providing a single method to infer regulatory programs for individual genes and gene modules. Although the idea of inferring networks from expression alone is attractive, genome-scale inference of such networks using current methods is an open challenge. In particular, the DREAM5 competition, which performed one of the largest comparative studies to date of network inference methods, showed that the agreement between the inferred networks and experimentally measured TF-target gene interactions is close to random in unicellular eukaryotes such as yeast [14].

One direction of work to address this issue is to impose additional constraints in the form of prior knowledge on the network structure to favor certain edges with auxiliary support from other data types, such as the presence of DNA-binding motifs for specific transcription factors. While the idea of incorporating priors into a probabilistic model, such as a Bayesian network [15], was proposed several years ago [16, 17], only recently has it been used on a genome-wide scale to integrate diverse types of evidence. Recent studies have implemented prior usage within a dependency network learning framework [18], where the network structure is inferred by solving a set of linear or non-linear regression functions: Greenfield et al.’s [19] extension of regularized linear regression in Modified Elastic Net (MEN) and Bayesian Best Subset Regression (BBSR), and iRafNet [20] (extending GENIE3 [21]). iRafNet incorporates different types of prior networks (protein-protein interactions, knockout data, gene expression time courses), whereas MEN and BBSR take a single input prior network.

Prior-based methods still rely on the predictive power of the mRNA level of the transcription factor, which might not reflect the cellular activity of a TF [22]. Towards this end, Arrieta-Ortiz and Hafemeister et al. [23] first use network component analysis [24] to predict the activity of TFs given a prior network, and then use these TF activities to predict a regulatory network for the bacteria B. subtilis. A large fraction (~62%) of the predicted interactions had experimental support indicating high accuracy in their predictions. While this shows early promise for prior-based methods for bacteria, an important future direction is to apply such approaches to complex eukaryotes (e.g. in plants and mammals). A key requirement of such approaches is prior knowledge of cis-regulatory elements on DNA sequence specificity of TFs. Experimental approaches to measure sequence specificity of TFs in vitro (e.g. using protein binding microarrays [25]) and in vivo (e.g. using DNase I (open chromatin) [26], ATAC-seq [27] and chromatin state [28]),
combined with computational modeling [29, 30], could significantly improve the quality of the prior networks for transcriptional regulatory networks.

More complete and accurate networks require measuring and modeling additional levels of transcriptional and post-transcriptional regulation. Accordingly, new methods are being developed to integrate these aspects. For example, using genome-wide miRNA and mRNA levels measurements from cancer samples, several groups have inferred miRNA-target gene networks [31, 32]. Osmanbeyoglu et al. [33] infer latent activity of and relationships between signaling proteins and transcription factors using a new regularized regression approach, integrating both reverse phase protein array levels and gene expression. Metabolic networks also have regulatory roles and interact with transcriptional networks to control the overall state of organisms [34, 35]. Recent availability of large scale metabolomic profiles have fueled parallel development of network-based methods to analyze metabolomic data. We refer the reader to recent reviews on network approaches with metabolomic data [36, 37] and integration of these networks with transcriptional networks [38, 39]. A direction that can be particularly useful for capturing different levels of regulation is to better model the distribution of the random variables representing the molecular entities, for example, by allowing heterogeneous random variables [40, 41].

Inference of a causal network by asserting directionality on edges is another important challenge in network reconstruction. Gene expression data from natural variation populations [42] or single gene perturbations [43] can help assert directed causal connections. To this end, the caPC (covariate-adjusted PC) algorithm [44] infers a partially directed graph by regressing each gene’s expression on a set of single-nucleotide polymorphisms (SNPs), and then uses the PC algorithm [45] to infer directionality based on conditional independence tests. Temporal information can also capture causal dependencies and dynamic Bayesian frameworks are particularly promising to address this [46]*. An alternative is to focus on inferring the directed connections among regulators only, using a prior regulatory network as input and inferring the activity of regulators rather than inferring new edges to targets. The biRte method [47] extends Nested Effects Models [48], to infer a directed graph over the regulators based on the identity and expression of their targets. An advantage of this method is that it does not predict regulator activity only based on mRNA levels. These approaches are promising steps toward solving a key challenge of inferring directionality in regulatory networks.

Context-specificty and dynamics in network reconstruction

The previous section focused on learning a single network from data. Often it is important to infer how networks change between different contexts (e.g., cell types and diseases) or over time. Recently, several methods have been developed to jointly infer networks that each represent a different context. Because information is shared during the inference process, the networks can be used to study shared and context-specific components.
The vast majority of methods have used Gaussian Graphical models (GGMs) [49, 50] that infer undirected “functional” interactions corresponding to direct statistical dependencies (Box 1C). Learning a GGM translates to estimating the non-zero off-diagonal entries in the inverse of the covariance matrix (also called the precision matrix), further imposing regularization terms to encourage sparsity. One approach by Kling et al. [51]* uses an augmented Sparse Inverse Covariance Selection (SICS) approach to jointly estimate one network (precision matrix) for each cancer, each network predicting connections between genes, microRNAs and sequence mutations. This is accomplished by a regularized regression framework with a novel multi-graph prior that encourages similarity between the networks as well as a modular structure within each network. Other approaches incorporate prior information about the relationship between the contexts. Treegl [52] and GNAT [53]* use a known hierarchy over contexts, such as a cell lineage. In Treegl, data are observed at both internal and leaf nodes, while for GNAT, data are available only in the leaf nodes.

Aiming to get closer to a directed regulatory network, Ontogenet [54]* infers per-module regulatory programs for multiple cell types in a lineage. This approach is similar to Treegl and GNAT, except that in Ontogenet the covariates of a gene are restricted to annotated regulators such as transcription factors and one estimates regulators for modules rather than individual genes, making this approach amenable to small sample size problems. GNAT and Ontogenet are of particular interest because they demonstrate that these approaches can be used on a large scale.

Context-specific networks have also been examined using supervised learning by training on a gold-standard network extracted from known interaction databases (e.g., [55, 56]) that is filtered based on node expression in a specific condition [57]. Park et al. [58] train tissue-specific support vector machines to predict the probability of an edge belonging to a particular interaction type (transcriptional regulation, phosphorylation, protein co-complexes, and post-translational regulation). Greene et al. [59]* apply a Bayesian data integration approach to learn tissue-specific networks for 144 tissues based on available human gene expression data sets in the Gene Expression Omnibus [60].

A more fine-grained notion of context is time; however, the above approaches are not applicable here because of the lack of sufficient samples per time point. Yosef et al. [61] overlayed time-point specific gene expression on a skeleton regulatory network inferred from sequence-specific motifs and targets. ODE-based models have the finest possible resolution [62], however, learning such models on a genome-scale is intractable and most approaches have focused on few dozen nodes [46]. To overcome the computational complexity while modeling different types of dependencies including temporal dependencies, the Jump3 method [63] uses tree-based models on time-series data and were able to model hundreds of genes.

Of special interest are methods that use single cell data such as single cell RNA-seq [64, 65] or proteomic levels [66] for inferring regulatory networks in cell fate specification problems [67, 68, 66]. The method of Ocone et al. [67]** uses dimensionality reduction techniques and cellular trajectory learning [69] to infer a pseudo time course.
They then use an ODE-based model to infer a regulatory network for each of the developmental branches. Because these assays can measure hundreds to thousands of cells, each cell providing a simultaneous measurement of multiple molecules at a time, these technologies open up a new avenue of inferring networks per cell type.

In summary, network reconstruction is the problem of inferring regulatory causal connections between molecular nodes in one or multiple conditions. Expression-based network inference methods have been popular because mRNA levels are the most widely available type of data. However, expression alone is not sufficient to learn regulatory networks, and new approaches that either incorporate prior interactions or additional types of data will be important to learn more accurate and complete networks.

Network-based interpretation and prioritization

Once a network structure is available, powerful computational approaches have been developed to tackle a variety of problems, broadly grouped under “network-based interpretation and prioritization”. Here we address three broad classes of methods: integrating gene sets from multiple high-throughput experiments, examining perturbation in networks, and prioritizing genes for follow up experiments. Common to these approaches is to use the network as a “skeleton” to define the universe of possible physical (e.g., protein-protein, protein-DNA) and functional (genetic) relationships among genes (Box 1C).

Integrating gene hits from complementary high-throughput experiments

Frequently, multiple high-throughput experimental assays such as transcriptomic or proteomic profiling, as well as large-scale functional screening studies [70, 71], are used to identify genes important for a specific biological process [72]. Often, the experimental methodologies are complementary and identify only partially overlapping hit gene sets. An important question is if, and how, the genes identified from each assay are related to each other. Network flow methods (Fig 1D) are popular approaches to identify subnetworks integrating input gene sets based on their connectivity on a skeleton network (reviewed by Kim et al. [73]). The key idea is that the subnetwork depicts how information flows between the different hit sets. For example, a subnetwork can link functional screen hits (upstream) to differential mRNA expression (downstream). Network flow methods often (but not exclusively) seek to find a minimal or sparse subnetwork (i.e., using a small number of edges to connect many hits) and vary in the node-, edge- and path-level properties that are inferred for each subnetwork. For example, prize-collecting Steiner trees [74] and maximum flow [72] methods define node- or edge-level properties (e.g., sign, directionality) and define an objective function that prefers a minimal network. In contrast, candidate path-based methods can incorporate more complex path-level properties [75, 76, 77, 78, 79], for example, enabling the connection of specific node pairs.
However, enumerating candidate paths or optimizing over them can be computationally challenging. As in network reconstruction, an emerging theme among information flow methods is to extend existing approaches to jointly infer subnetworks for multiple related contexts [79, 80, 81].

All methods define these network level properties as constraints and solve the constrained optimization method using either: (i) linear (or integer linear) programming (ILP) methods [72, 74, 76, 77, 78, 79], or (ii) statistical inference within a maximum likelihood framework [75, 82, 83]. When considering genome-scale networks, ILP-based methods have the advantage in that they can find globally optimal solutions more efficiently compared to probabilistic approaches. However, probabilistic approaches are also useful at smaller scales as different parameters can be estimated in a data-driven manner.

**Interpreting genetic variation**

A network also provides molecular context to interpret genetic variation (Fig 1D), including Single Nucleotide Polymorphisms (SNPs). Some approaches are applicable when there is a known directionality from the perturbation to specific affected targets. **PINE (Perturbations in NEtworks)** [84] is a probabilistic method to predict which branches of an input parameterized signaling network are affected by genetic variants under specific conditions. If a less detailed input network is sufficient, network flow methods (e.g., *minimal set cover*) can also be used. For example, in cancer, identifying causal “driver” variants from correlated “passenger” mutations is important. This problem is commonly addressed by *minimum set cover* approach (recently, [85, 86]) to distinguish causal driver mutations from passenger' mutations by finding a small set of drivers that can be connected to the targets by paths in a protein-protein network.

If only a list of mutations is available, or if causal directionality from mutations to gene expression is not assumed, one can use *module-based methods* to find densely connected components of the network that are enriched for highly mutated genes (e.g., [87, 88]). Grouping the mutated or dysregulated genes into modules may make the approach more resistant to noise in the network or data. Recently this concept has been applied in conjunction with expression data to interpret genetic variation in autism [89] and glioblastoma [90].

A limitation of current skeleton-network based approaches is that they typically include only perturbations to coding regions, and cannot be used to interpret regulatory variants, many of which can lie far away from genes [91]. As our ability to identify regulatory variation improves, for example from high-throughput regulatory genomic datasets [6, 5], extending network-based methods to incorporate regulatory variants will become increasingly important [92, 93].
**Gene prioritization**

Protein-protein and functional interaction networks are often used to prioritize genes for further experiments (Fig 1E). Most of these approaches (reviewed by Moreau & Tranchevant [94]) have either relied on network centrality measures (e.g., connectivity of hub genes, Box 1D, [95]), or, have used an input set of known relevant genes to rank candidates using the concept of guilt by association: genes that are close to input genes (e.g., based on shortest path length, or more global diffusion-based measures [96]) in the skeleton network are likely to have similar function. An increasingly important and exciting new direction is to prioritize genes to efficiently convert one cell type into another, e.g., to reprogram differentiated cells into induced pluripotent cells [97, 98]. **CellNet** [98]** first infers tissue and cell type-specific networks, and then uses those networks in a Random Forest classification approach to prioritize genes for reprogramming and to quantify similarity between a query sample and a known cell type. A subsequent approach, **Mogrify** [97], prioritizes genes based on proximity on an input skeleton network and a measure of differential expression, which is informed by the lineage structure of cell types. Both approaches successfully recapitulated known cell-fate specification genes, and also predicted new genes that were validated to be important regulators for establishing a specific fate.

In summary, network-based methods can be powerful for interpretation of high-throughput measurements and for prioritization of experiments. The choice of the skeleton network is important as most approaches rely on interactions from public databases that may not be relevant to the cellular context being modeled. Recent efforts to infer context-specific interactions may allow these network-based approaches to achieve higher precision [59].

**Conclusions**

Network-based computational methods aim to build, interpret and integrate molecular networks and have been important tools for studies of complex responses. In this review, we discussed recent promising approaches for two major problems: network reconstruction and network-based interpretation. Going forward we envision addressing one major challenge: to build network-based causal, predictive models of complex phenotypes. This will require us to construct networks and prioritize experiments in concert, rather than independently, within an integrated computational-experimental framework that permits multiple iterations of prediction and validation. Novel network-based models that leverage new types of datasets from emerging technologies for measuring RNA [65] and open chromatin [99] for single cells, as well as approaches to perform targeted perturbations (e.g. using CRISPR/Cas9 technologies [100, 101]) will increasingly enable us to infer causal edges and identify the mechanistic, molecular underpinnings of complex traits and diseases.
Acknowledgments

SR is supported by a National Science Foundation CAREER award (NSF DBI: 1350677) and a Sloan Foundation research fellowship. This publication was made possible in part by US Environmental Protection Agency grant 83573701.
Annotated bibliography

The following papers are of special (*) or outstanding (**) interest:

Osmanbeyoglu et al, 2014 [33]*: A novel regularized regression framework (affinity regression) for inferring relationships between signaling proteins and transcription factors using protein and mRNA levels. Notably, this approach can make predictions for both types of data. Given a new protein (or gene) expression profile, it can be used to predict active transcription factors (or signaling proteins).


Ocone et al, 2015 [67]**: Using the idea that individual cells in a single sample are at different stages in a developmental or response process, the authors present an approach for learning a dynamic regulatory network from single-cell RNA-seq data.

Kling et al, 2015 [51]*: Method for integrating multiple genomic measurements across many sample cohorts by jointly estimating regularized inverse covariance matrices.

Pierson et al, 2015 (GNAT) [53]* Transfer-learning approach to identify context-specific co-expression networks given a hierarchy over contexts. Application to 35 GTEx tissues identified tissue-specific components that could not be explained due to differences in nodes.

Jojic et al, 2013 (Ontogenet) [54]*: A multi-task learning approach for inferring modular regulatory programs for 195 cell types in a developmental lineage. Importantly, the approach requires only a few samples per cell type.

Greene et al, 2015 (GIANT) [59]* Large-scale prediction of tissue-specific interaction networks using massive amounts of publicly available microarray data and a network-based prioritization tool for sequence variants. The authors provide a web interface (GIANT) for their networks and prediction tools.

Cahan et al, 2014 (CellNet) [98]** Network biology approach for inferring cell type and tissue-specific networks, as well as prioritizing regulators for directed reprogramming of stem cells. Validated by experiments that demonstrate its use for improving reprogramming efficiency.

Bibliography


8. GTEx Consortium. The genotype-tissue expression (gtex) project. *Nat Genet* 2013;45(6):580–585. URL: http://dx.doi.org/10.1038/ng.2653. doi:10.1038/ng.2653.


Network terminology. A molecular network is defined by the type of entities it connects and the information it contains about those connections.

A. Network elements. Nodes in a molecular network are intracellular molecules, frequently genes or proteins. Edges represent interactions (binding, regulatory, post-translational modification, correlation) and may be directed (if it is known that the first molecule’s activity or function influences the second’s) or undirected.

B. Network structure and function. On the right is a cartoon showing a kinase $g_1$ and transcription factor (TF) $g_2$ whose coordinate activity regulates a gene $g_3$. The structure shows $g_1$ and $g_2$ as regulators of $g_3$, and the function $f$ quantifies the relationship between regulator state and target. This function estimates the state or activity level of the target gene in one or more than time points/conditions as a function of state/activity of regulators, or other input parameters. In the literature, this function has taken many forms, including simple linear regression, linear or non-linear differential equation, and probabilistic Boolean function.

C. Types of networks. Regulatory networks describe transcriptional regulatory programs (as in Panel B). Signaling networks describe cascades of post-translational modifications from signaling receptors to transcription factors. Functional networks are undirected, and represent statistical dependencies between nodes. Physical networks represent experimentally observed binding interactions between proteins or between proteins and DNA or RNA. In any type of network, weights on edges may correspond to statistical significance, confidence, or inferred importance of the edge.

D. Organizational units. Networks are composed of smaller structural units. A subnetwork is a subset of nodes and edges from the original network; subnetworks are used to predict functional coordination between nodes. A regulatory path is a chain of nodes and edges that posits the ordered mechanism by which upstream regulators and transcription factors influence the regulation of a target gene. A regulatory module is a group of genes that are co-expressed and are predicted to share the same regulatory program. A hub node has a high degree (number of neighbors); hubs are often predicted to be important regulators of one or more cellular pathways.

Box 1: Network terminology

Figure 1: Overview of major computational network biology tasks addressed here. Shown are the key inputs, algorithmic approaches and outputs produced in each type of task. (A) Network reconstruction for identifying a single network or multiple context-specific networks. (B) Network-based interpretation to identify a relevant subnetwork that integrates diverse data types (e.g. mRNA and proteomic measurements) on a skeleton network. (C) Network-based interpretation of genetic variants to identify either subnetworks or network modules. (D) Network-based prioritization to prioritize genes for further experiments based on an input gene set and measured expression levels and proximity of candidate genes on the skeleton graph.


### Problem

<table>
<thead>
<tr>
<th>A. Reconstruct networks (one or more contexts)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
</tr>
<tr>
<td>Expression levels</td>
</tr>
<tr>
<td>mRNA proteins (some methods)</td>
</tr>
<tr>
<td>Prior network (some methods) (motif, ChIP, PPI)</td>
</tr>
<tr>
<td><strong>Do</strong></td>
</tr>
<tr>
<td>• Sparse regression</td>
</tr>
<tr>
<td>• Random Forests</td>
</tr>
<tr>
<td>• Graphical models</td>
</tr>
<tr>
<td>• Causal networks</td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
<tr>
<td>Regulatory network</td>
</tr>
<tr>
<td>Structure</td>
</tr>
<tr>
<td>Regulators</td>
</tr>
<tr>
<td>Target</td>
</tr>
<tr>
<td>Function</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Network-based interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
</tr>
<tr>
<td>Measurements</td>
</tr>
<tr>
<td>mRNA Protein</td>
</tr>
<tr>
<td>Skeleton network</td>
</tr>
<tr>
<td>Protein mRNA</td>
</tr>
<tr>
<td><strong>Do</strong></td>
</tr>
<tr>
<td>• Steiner trees</td>
</tr>
<tr>
<td>• Maximum flow</td>
</tr>
<tr>
<td>• Path-based methods</td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
<tr>
<td>Relevant subnetwork</td>
</tr>
<tr>
<td>Association</td>
</tr>
<tr>
<td>low high</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. Interpret genetic variants in living systems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
</tr>
<tr>
<td>Skeleton network and measurements</td>
</tr>
<tr>
<td>Protein mRNA Variant Dysregulated</td>
</tr>
<tr>
<td><strong>Do</strong></td>
</tr>
<tr>
<td>• Test associations between candidate variants, condition, targets</td>
</tr>
<tr>
<td>• Information flow</td>
</tr>
<tr>
<td>• Module finding</td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
<tr>
<td>Functional mutations, affected targets</td>
</tr>
<tr>
<td>Dense modules</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D. Prioritize genes for experiments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Given</strong></td>
</tr>
<tr>
<td>Any network</td>
</tr>
<tr>
<td>Input gene set</td>
</tr>
<tr>
<td><strong>Do</strong></td>
</tr>
<tr>
<td>Score genes:</td>
</tr>
<tr>
<td>• Distance to all input genes</td>
</tr>
<tr>
<td>• Expected expression outcome</td>
</tr>
<tr>
<td><strong>Output</strong></td>
</tr>
<tr>
<td>Scores for genes</td>
</tr>
<tr>
<td>Confidence</td>
</tr>
<tr>
<td>low high</td>
</tr>
</tbody>
</table>
Box 1

A. Node (gene, protein)
   - Directed edge
   - Undirected edge

B. Kinase $g_1$ → TF $g_3$
   - $g_3 = f(g_1, g_2)$
   - Structure $\implies$ Function

C. Undirected
   - Functional
   - Physical
   - Directed
   - Regulatory
   - Signaling

D. Subnetwork
   - Upstream regulator
   - Transcription factor
   - Target
   - Regulatory path
   - Hub node (many neighbors)
   - Regulatory module (targets in box)
graphical abstract

Measure
- Intracellular molecules (mRNA, protein, metabolite, ...)
- Interactions (protein-protein, protein-DNA, ...)

Infer networks
- Network for one condition
- Context-specific network

Use networks for interpretation
- Prioritize genes
- Interpret network perturbations
  - Affected genes
  - Perturbed genes
  - Priority low, high

Algorithms