PLOS | COMPUTATIONAL BIOLOGY

Education

Chapter 5: Network Biology Approach to Complex Diseases

Dong-Yeon Cho[9], Yoo-Ah Kim[9], Teresa M. Przytycka*

National Center for Biotechnology Information, NLM, NIH, Bethesda, Maryland, United States of America

Abstract: Complex diseases are caused by a combination of genetic and environmental factors. Uncovering the molecular pathways through which genetic factors affect a phenotype is always difficult, but in the case of complex diseases this is further complicated since genetic factors in affected individuals might be different. In recent years, systems biology approaches and, more specifically, network based approaches emerged as powerful tools for studying complex diseases. These approaches are often built on the knowledge of physical or functional interactions between molecules which are usually represented as an interaction network. An interaction network not only reports the binary relationships between individual nodes but also encodes hidden higher level organization of cellular communication. Computational biologists were challenged with the task of uncovering this organization and utilizing it for the understanding of disease complexity, which prompted rich and diverse algorithmic approaches to be proposed. We start this chapter with a description of the general characteristics of complex diseases followed by a brief introduction to physical and functional networks. Next we will show how these networks are used to leverage genotype, gene expression, and other types of data to identify dysregulated pathways, infer the relationships between genotype and phenotype, and explain disease heterogeneity. We group the methods by common underlying principles and first provide a high level description of the principles followed by more specific examples. We hope that this chapter will give readers an appreciation for the wealth of algorithmic techniques that have been developed for the purpose of studying complex diseases as well as insight into their strengths and limitations.

This article is part of the "Translational Bioinformatics" collection for *PLOS Computational Biology*.

1. Introduction

Complex diseases are caused, among other factors, by a combination of genetic perturbations. Thus in the case of a complex disease we do not assume that a single genetic mutation can be pinned down as a cause. Many diseases fall in this category including cancer, autism, diabetes, obesity, and coronary artery disease. Even though there are other factors involved in such diseases, this review will focus on genetic causes.

One of the fundamental difficulties in studying genetic causes of complex diseases is that different disease cases might be caused by different genetic perturbations. In addition, if a disease is caused by a combinatorial effect of many mutations, the individual effects of each mutation might be small and thus hard to discover. For example, autism is considered to be one of the most heritable complex disorders, but its underlying genetic causes are still largely unknown [1]. One of the proposed factors that contribute to this difficulty is the role of rare genetic variations in the emergence of the disease [2].

An additional difficulty in studying complex diseases relates to disease heterogeneity. Specifically, in a complex disease, disease phenotypes might vary significantly among patients. The recognition of this fact has lead, for example, to renaming "autism" to "autism spectrum disorders" (ASDs) referring in this way to a group of conditions characterized by impairments in reciprocal social interaction and communication, and the presence of restricted and repetitive behaviors [1]. Similar heterogeneity is present in other complex diseases including cancer.

Given the above challenges, how can we approach the study of complex diseases? A useful clue is provided by the fact that genes, gene products, and small molecules interact with each other to form a complex interaction network. Thus a perturbation in one gene can be propagated through the interactions, and affect other genes in the network. However, the fact that we observe similar disease phenotypes despite different genetic causes suggests that these different causes are not unrelated but rather dys-regulate the same component of the cellular system [3]. Therefore in studies of complex diseases researchers increasingly focus on groups of related/interconnected genes, referred to as modules or subnetworks.

2. Interactome

Biomoecules in a living organisim rarely act individually. Instead, they work together in a cooperative way to provide specific functions. A variety of intermolecular interactions including protein-protein interactions, protein-DNA interactions, and RNA interactions are essential to these cooperative activities. These interactions can be conveniently represented as networks (graphs) with nodes (vertices) which denote molecules, and links (edges) which denote interactions between them. Depending on the type of interaction, the corresponding edge might be directed or

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: przytyck@ncbi.nlm.nih.gov

[9] These authors contributed equally to this work.

undirected. For example, a binding between two proteins is usually represented as an undirected edge while an interaction between a transcription factor and a gene whose expression is regulated by the given transcription factor is usually represented as a directed edge where the direction goes from the transcription factor to the gene.

Biological interaction networks have characteristic topological properties [4]. One of the basic properties observed in many biological networks is the scale-free property [5]. A scale free network is defined as a network whose node degree distribution follows a power law. Formally, the function $P(k)$ indicating the fraction of nodes interacting with k other nodes in the network follows $P(k) \sim \alpha k^{-\gamma}$, where α is a normalization constant and the degree exponent γ is usually in the range of $2 < \gamma < 3$. Obviously, in biological networks the scale free property holds only approximately and practically the most important implication of this observation is the fact that these networks are characterized by a small number of highly connected nodes while most nodes interact with only a few neighbors. These highly connected nodes, called hubs have been proposed to play important roles in biological processes [6] and shown to be related to the modular structure of the physical and functional interaction networks [7]. Therefore it might be interesting to consider disease related genes in the context of the topological properties of interaction networks such as connectivity or modularity [8,9]. With respect to connectivity, one should note that known disease genes tend to be more studied which might introduce a bias towards higher connectivity. Importantly, independently of the source of the non-uniformity of node degree distribution, this characteristic property of interaction networks needs to be kept in mind while designing proper null models for conclusions derived using these networks.

In the following subsections, we briefly describe how physical and functional interactions networks are constructed and how they are applied to analyze complex diseases. We then explore the modularity of networks – a widely accepted phenomenon in biological networks that has proven to be helpful in disease studies.

2.1 Physical Interaction Networks

Physical contacts between proteins are critical in many biological functions. In fact much of the molecular machinery responsible for transcription, translation, and degradation is made of stable protein complexes. There are two main approaches for detecting physical protein interactions [10]. The first approach is to detect physical interactions between protein pairs. The most widely used high-throughput technology for detecting pairwise interaction is yeast two-hybrid (Y2H) method. Alternatively, physical interactions among groups of proteins can be detected without explicit consideration of interacting partners. For this type of approach, interaction data is typically obtained by tandem affinity purification coupled to mass spectrometry (TAP-MS). A more detailed review on experimental methods for the detection and analysis of protein-protein interactions can be found in [11]. It is worth noting that networks obtained with various technologies often have different topological properties [7]. For example, in the case of the yeast TAP-MS network, hub nodes are enriched with essential genes (the genes without which yeast cannot survive in standard growth medium). In contrast, hubs in yeast Y2H networks are enriched with genes that are pleiotropic [12]. Finally, experimental procedures detecting protein-protein interactions have also been complemented by various computational methods using evolutionary-based approaches, statistical analysis, and/or machine learning techniques (for a review, see [13]).

While these physical interaction networks have significantly advanced our understanding of the relationships between molecules, a concern is their level of noise and incompleteness. Indeed, physical interaction networks obtained by high-throughput techniques are found to include numerous non-functional protein-protein interactions [14] and at the same time many missing true interactions. Therefore physical interactions are often complemented with functional interactions.

2.2 Functional Interaction Networks

While physical interaction networks provide information on how proteins interact with each other, sometimes we may be more interested in how proteins work together to perform a certain function. Functional networks aim to connect genes with similar or related functions even if they do not necessarily physically interact. Similarly functional regulatory networks are constructed so that the interactions depict direct or indirect regulatory relationships. Consequently, several computational methods have been proposed to derive functional interaction networks.

Since functionally related genes are likely to show mutual dependence in their expression patterns [15], gene expression data has been often used to detect functional relationships. Co-expression networks can be constructed by computing correlation coefficients or mutual information between gene expression profiles of every pair of genes in different experimental settings. To build more comprehensive functional networks, co-expression data is frequently combined with other types of data such as Gene Ontology [16,17], outcome of genetic interaction experiments, and physical interactions. Such integrated networks have been constructed for a variety of organisms including yeast [18], fly [19], mouse [20], and human [21].

Gene regulatory network reconstruction algorithms such as ARACNE [22] and SPACE [23] identify regulatory relationships building on the assumption that changes in the expression level of a transcription factor should be mirrored in the expression changes of the genes regulated by the transcription factor (TF). Causal relations among genes can also be naturally modeled using Bayesian networks which can represent conditional dependencies between expression levels (for a primer on Bayesian network analysis utilizing expression data (see [24]); for a recent review see [25]). Considering the temporal aspects of gene expression profiles, dynamic

Bayesian networks have been used to model feedback loops as well as gene regulation patterns [26,27]. While expression profiles serve as primary data sources for constructing functional regulatory networks, this data is often complemented with additional information such as experimentally derived transcription factor binding data from ChiP-seq experiments or computationally identified binding motifs.

2.3 Modules and Pathways

It is widely accepted that the cellular system is modular. Hartwell *et al.* defined a functional module as an entity, composed of many types of interacting molecules, whose function is separable from those of other modules [28]. While the precise meaning of separation is left undefined, this general description provides a good intuition behind the concept of a module. Traditionally, molecular pathways have been delineated by focused studies of particular functions such as cell growth. Typically, these pathways contain not only topological connectivity information but also the roles of molecules such as whether a given molecule is an activator or inhibitor of the activity of another molecule. However, these hand-curated pathways are often incomplete. In addition, while some functions, such as cell growth or differentaition, have been relatively well studied, studies of other pathways are less extensive. Therefore, given the availability of large scale interaction networks, it is natural to attempt to extract meaningful functional modules from such networks. While there is no unique way to mathematically define functional modules, the most common approach is to search for densely connected subgraphs or clusters [29–46]. Additionally, gene expression information can be used alone or in concert with protein interaction data to obtain gene modules by grouping co-expressed genes into one module [47–49].

It is important to keep in mind that modules identified by analysis of high-throughput data are noisy, containing both false negative and false positive edges. In addition they do not usually provide information about the nature of an interaction. Therefore, unlike hand curated pathways, computationally identified network modules typically lack a mechanistic explanation of pathway activities but rather serve as groups of genes that work together to achieve a particular function.

An important advantage of working with modules rather than individual genes relates to the fact that it is often easier to predict the function of a module than the function of a gene. In particular, while the functions of many genes are still unknown, the prediction of the functional role of a module may be possible if the module contains a sufficient number of genes of known functions. Such enrichment analysis builds on the assumption that a fraction of genes can be assigned a functional category such as Gene Otology (GO) term [17]. The question of whether the number of genes with a functional annotation in a given gene module is higher than expected by chance can be determined by statistical tests such as χ^2 or Fisher exact test. A variety of software tools have been developed to perform such an analysis [50].

3. Identifying Modules and Pathways Dys-regulated in Diseases

Since complex diseases are believed to be caused by combinations of genetic alterations affecting a common component of the cellular system, module-centirc approaches are particularly pormissing in thier study. How can disease associated modules/subnetworks be identified? Complementing interaction data with additional data related to disease states helps in separating subnetworks perturbed in a disease of interest from the remainder of the network. Both genotypic data (e.g., SNP, copy number alteration) and molecular phenotypic data such as gene expression profiles in disease samples have been used to aid the identification of perturbed network modules and explain the connection between genotypic and phenotypic data (reviewed in [51]). Basing on the assumption that complex diseases are caused by a set of mutations which, although strongly vary among patients, are likely to dys-regulate common pathways, such dys-regulated pathways might be uncovered by mapping genes altered in the diseases onto a PPI (protein-protein interaction) network and then searching for network modules enriched with the altered genes. On the other hand, organismal level phenotypes such as diseases are directly related to molecular level changes such as gene expression. Thus an alternative group of approaches considers modules enriched with abnormally expressed genes. Finally, molecular pathways can also be considered as means of information flow. For example, the activation of the EGFR signaling pathway starts with the activation of the EGFR receptor, which in turn activates a number of signaling proteins downstream which initiate several signal transduction cascades, such the MAPK, Akt and JNK pathways and culminate in cell proliferation. Thus

the third type of approaches focuses on predicting molecules and modules that mediate such information propagation.

What are the benefits of analyzing phenotypic and genotypic differences in diseases in the context of their molecular interactions? First, the integrative network based approaches can identify subnetworks that include genes that do not necessarily show a significantly different state in disease versus control but still play an important role within a module by mediating a connection between other disease associated genes. For example, in their pioneering approach, Ideker *et al.* [52] integrated yeast protein–protein and protein-DNA interactions with gene expression changes in response to perturbations of the yeast galactose utilization pathway and identified *Active Subnetworks* (sets of connected genes with significantly differential expression) which included common transcription factors showing moderate changes in their gene expression level but connecting other dys-regulated genes. Second, a module based approach increases statistical power, allowing the identification of a perturbed module even in the case when the perturbation of each individual gene in the module might not be statistically significant. For example, many cases of genetic diseases such as autism and schizophrenia are affected by rare germline variations which are difficult to distinguish from noise due to their rarity. However, recent studies showed that a significant portion of the altered genes belong to a highly interconnected protein network [53], suggesting the network approach can better detect the causal genes. Third, identified network modules can provide better understanding of the biological underpinning of the diseases and therefore more reliable markers in disease diagnosis and treatments (see Section 4 for more discussion).

3.1 Network Modules Enriched with Genetic Alterations

One way in which differing genetic variations might dys-regulate a common pathway is when the genes containing these alterations belong to the pathway. This potential explanation has led to the idea that the dys-regulated pathways might be uncovered by mapping the genes altered in the diseases to an interaction network and searching for the modules enriched with the altered genes (See Figure 1).

Following this principle, the first step to identify such modules is to select candidate genes whose alterations may have caused a disease of interest. Genes or whole geno-
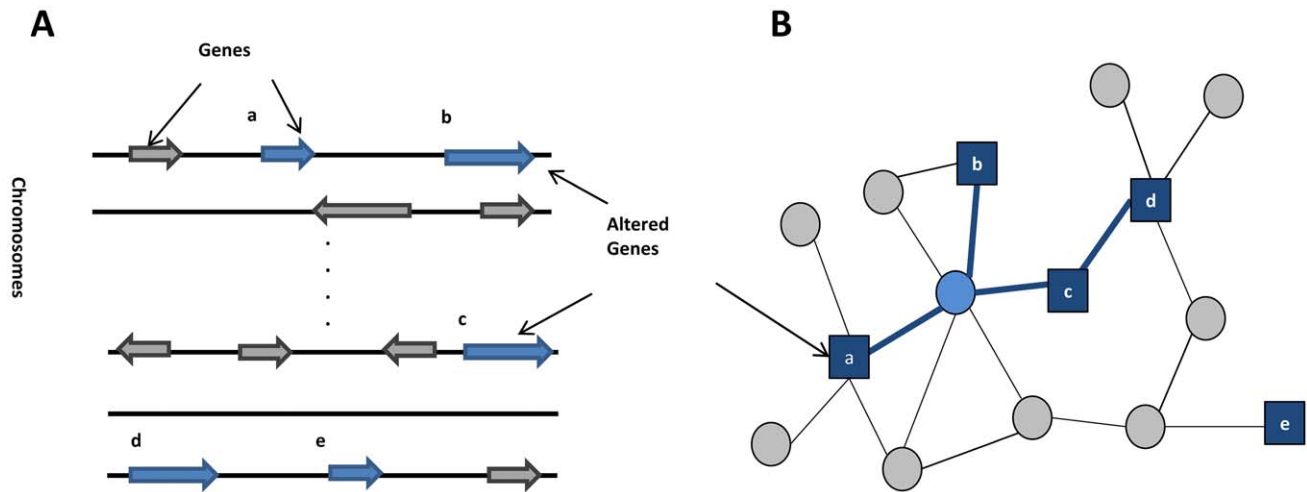
Figure 1. Identification of network modules enriched with genetic alterations. (A) Genomic regions with alterations. (B) Genes in the altered regions are mapped to the interaction network and modules enriched with such genes are identified.
doi:10.1371/journal.pcbi.1002820.g001

mic regions that are altered in the disease are first identified, and the genes residing in the altered regions are mapped to an interaction network. Both physical and functional interaction networks can be used, and edges might be weighted based, for example, on the likelihood of having the same phenotypes or influences between genes [54–56]. Next, modules are typically defined as subsets of genetically altered genes that are highly interconnected or within close proximity to each other in the interaction network together with non-altered genes necessary to mediate these connections. Edge weights, if given, can be used to prioritize the modules. In many cases, finding the best subnetwork is computationally expensive and search algorithms such as greedy growth heuristics or more sophisticated approximation algorithms have been proposed. Finally, rigorous statistical tests have been applied to evaluate the significance of selected modules.

Examples. The idea of finding genetically altered network modules has been utilized in various disease studies. Analyzing ovarian cancer TCGA data (The Cancer Genome Atlas), HOTNET identified subnetworks in a protein interaction network in which genes are mutated in a significant number of patients [54]. The identified networks includes the NOTCH signaling pathway which is indeed known to be significantly mutated in cancer samples [57]. The method is based on the set cover approach (see Set cover based approach section below), which is found to be effective in capturing different genetic variations across patients. In the NETBAG (NETwork-Based Analysis of Genetic associations) method,

developed by Gilman *et al.* and applied to identify a biological subnetwork affected by rare de novo copy number variations (CNVs) in autism [58,59], the authors first constructed a gene network where edges were assigned the likelihood odd ratio for contributing to the same genetic phenotype. Subsequently a greedy growth algorithm was used to find clusters in this network. In another approach, Rossin *et al.* [60] considered the genomic regions found to be associated with Rheumatoid Arthritis (RA) and Crohn's disease (CD) in previous GWAS studies, and connected the genes residing in these regions based on interaction data to obtain network modules. It was also verified that those identified modules exhibited significant differences in expression level in the disease samples.
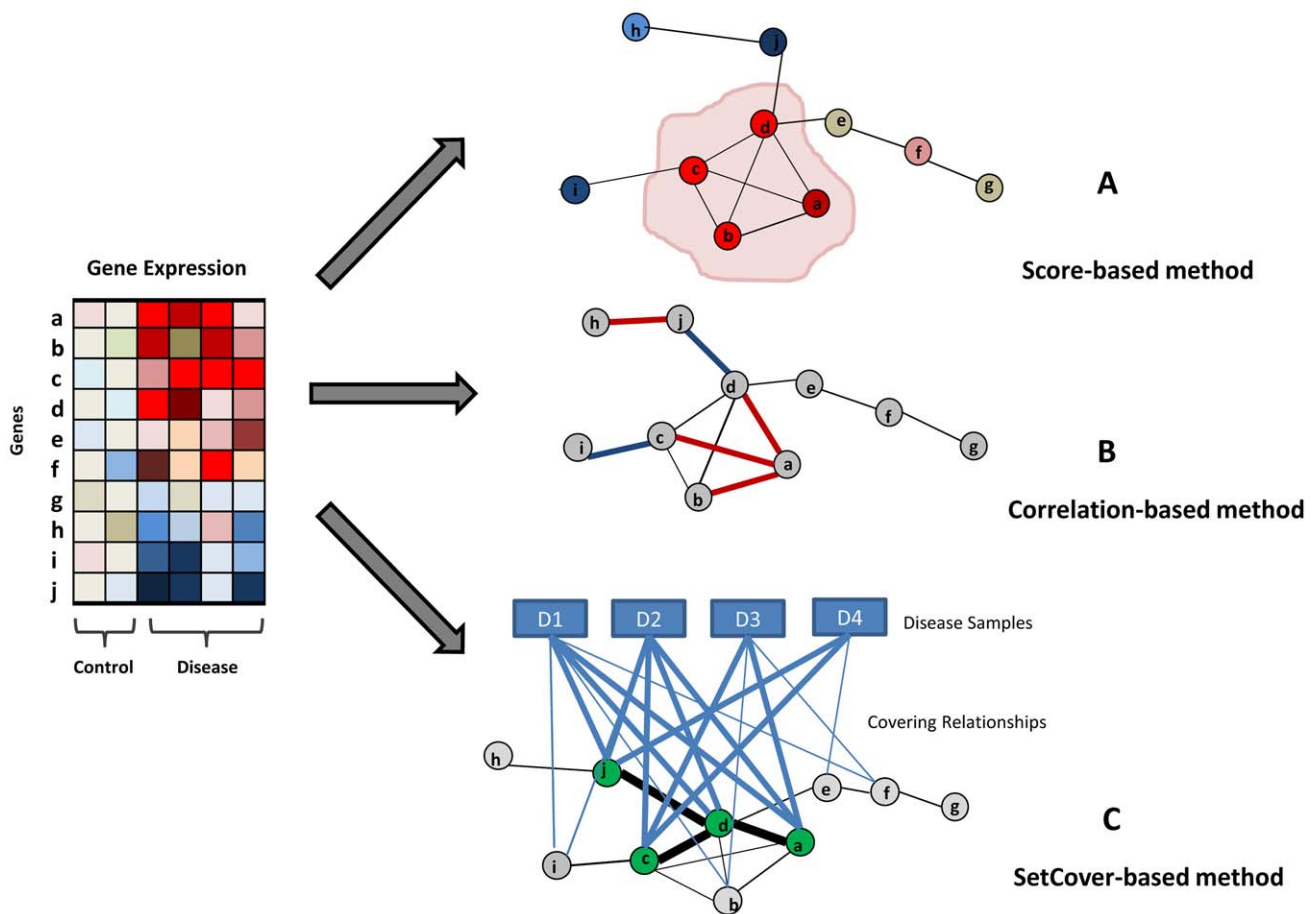
3.2 Differentially Expressed Network Modules

Another popular and successful approach to find disease associated modules is to search for subnetworks that are significantly enriched with genes whose expression levels are changed in disease samples. Building on the observation that a molecular perturbation typically affects the expression levels of genes in a whole module rather than individual genes, these approaches identify the modules which exhibit different expression patterns in disease states relative to a control. Gene expression data has been widely utilized for identifying dys-regulated modules and drug targets, inferring interactions between genes, and classifying diseases. While these approaches are based on the common idea of finding gene modules

enriched with genes that have abnormal expression, several different computational techniques have been used to achieve these tasks, which we discuss shortly below. The methods are also illustrated in Figure 2.

3.2.1 Scoring based methods. Suppose that there is a subset of genes which are differentially expressed in disease samples and they are closely connected to each other in an interaction network. A subnetwork including such genes might be a good candidate for a disease associated network module (Figure 2A). Implementing this idea requires a way to score candidate modules. Various methods have been suggested for measuring the significance of the differential expression of genes in a module and their connectivity (the distances between the genes). In addition, different methods adopt different search algorithms to find high scoring candidate modules. Finally, some approaches additionally require that all genes are either up-regulated or down-regulated in the same direction.

Examples. Chuang *et al.* defined the activity score for a subnetwork by comparing gene expression profiles from two different types of samples (metastatic or non-metastatic in their study) [61]. More specifically, they first computed how well the expression of a gene discriminates between the two patient groups and then scored candidate subnetworks based on aggregate discriminative power over all genes in the subnetwork. Then they searched for the most discriminative networks in a greedy manner. While the method was used for disease classification

Figure 2. Finding differentially expressed modules. (A) Score based method selects the module with significant expression changes. (B) Correlation based method selects edges with correlation changes. The red and blue edges are correlated and anti-correlated edges, respectively. (C) Set cover based method selects a set of genes covering all samples. In this example, each sample has at least 2 differentially expressed genes and the genes are connected in the network.
doi:10.1371/journal.pcbi.1002820.g002

(see Section 4), it can readily be applied to leverage the difference between disease and non-disease cohorts.

3.2.2 Correlation based methods. Comparing expression patterns between genes is a basis for constructing a co-expression network, extracting modules exhibiting similar expression patterns, and further understanding molecular changes in diseases. Considering expression correlation of disease cases in the context of interactions can provide additional power in the identification of a disease associated module (Figure 2B). If the expression changes of two neighboring nodes are correlated with each other, this may suggest that the two interacting genes have related functional roles. With this in mind, some approaches look at connected components which show highly correlated and anti-correlated expression patterns. Other approaches search for loss and gain of correlation in disease states to identify dys-regulated edges.

Examples. Aiming to identify regulatory networks defining phenotypic classes of human cell lines, Müller *et al.* searched for Jointly Active Connected Subnetworks (connected subnetworks with high average internal expression similarity) in a human interaction network [62] and demonstrated the power of combining network and expression data.

IDEA (Interactome Dysregulation Enrichment Analysis) method [63] focused on the identification of perturbed network edges in a combined interaction network (PPI, transitional, signaling, posttranslational modifications predicted by MINDy [64]), and searched for the edges connecting genes which in a disease state show loss or gain of expression correlation. The utility of the method was demonstrated in the analysis of FL lymphoma and other cancer types. In particular, they identified BCL2 as the gene adjacent to the largest number of dys-regulated edges in FL lymphoma. This analysis also identified

the SMAD1 gene, which could not be detected by differential expression analysis only.

To understand the mechanism of aging, Xue *et al.* applied a network module approach [65,66]. They utilized a PPI network and overlaid expression data obtained from various stages of aging. Two types of edges – correlated and anti-correlated – were selected. The subnetwork that includes only those edges was called the NP (negative and positive) network, is proposed to be related to the aging mechanism. Further modularizing the network with hierarchical clustering of expression patterns, they obtained a few age related modules and found some genes connecting different modules through PPIs are more likely to affect aging/longevity, which was also experimentally validated.

3.2.3. Set cover based methods. A group of methods employ a combinatorial approach named set cover. In a set cover, a gene is considered to cover a disease
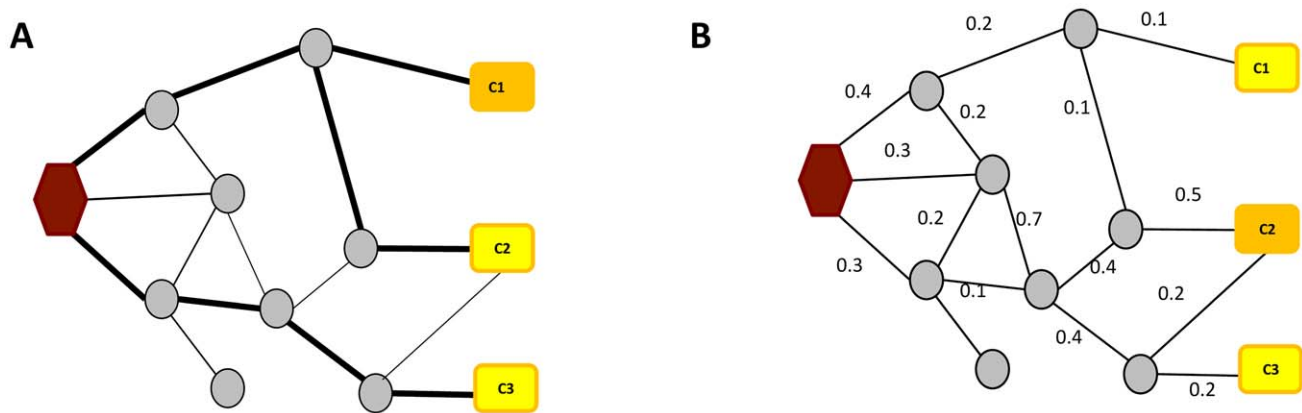
Figure 3. Finding information propagation modules. (A) Shortest path approach to uncover information propagation. The shortest paths from a target gene (with hexagon shape) to each of three candidate genes are shown. The closest gene is identified as the most probable disease causing gene. (B) Flow based approach. The gene receiving the most significant amount of flow is identified as the disease gene. The information flow methods often follow Kirchhoff's current law (the amount of incoming information equals the amount of outgoing information).
doi:10.1371/journal.pcbi.1002820.g003

sample if it is dys-regulated in the sample. For example, it can be decided if a gene is covering a sample or not based on the fold change of gene expression level in the sample or using a statistical test such as z-test. The main principle of the set cover approach is that each disease case has some dys-regulated (thus covering) genes but in heterogeneous diseases, different cases will typically have different covering genes. Set cover approaches provide a strategy to select a representative set of such covering genes (Figure 2C). This is usually done by defining some optimization criterion and attempting to select a set of genes which is optimal with respect to this criterion. For example, given a set of genes and disease samples along with covering relationships, a subset of genes is selected so that each sample is covered by some minimal number of genes while the total number of selected genes is minimized.

Many observed organism-level phenotypes arise in a heterogeneous way. Diseases such as cancer are now seen as a spectrum of related disorders that manifest themselves in a similar fashion. Since different samples may be covered by different genes and those genes may be connected in an interaction network, set cover approaches can be useful to identify gene modules explaining a heterogeneous set of samples [67–69].

Examples. Aiming to detect dys-regulated pathways in complex diseases, Ulitksy *et al.* extended the set cover technique by integrating expression data and interaction networks [67]. Their method, named DEGAS (de novo discovery of dys-regulated pathways) searches for a smallest set of genes forming a connected subnetwork

so that each disease sample is covered by a certain minimal number of genes. They applied this approach to a Parkinson's disease dataset. Chowdhury *et al.* [68], developed an alternative network cover based algorithm and used the identified modules for disease classification in a human colorectal cancer dataset.

Set Cover approaches have also been applied to data types other than gene expression. For example, Kim *et al.* proposed a module cover approach to identify gene modules which collectively cover disease samples [70]. At the same time they required that each module is coherent, containing genes with similar genotype-phenotype mappings (see Section 4 for more discussion). The HotNet Algorithm discussed in Section 3.1 also utilized a variant of a set cover approach to find genetically altered modules. In their case, a gene is defined to cover a sample if the gene is mutated in the sample, and they looked for a fixed size connected set of genes covering as many samples as possible. The Dendrix (De novo Driver Exclusivity) algorithm was also developed to discover mutated gene modules in cancer and, though it does not utilize interaction data, it aims to find sets of genes, domains, or nucleotides whose mutations exhibit both high coverage and high exclusivity in the disease samples [71].

3.3 Uncovering Information Propagation Modules

The approaches discussed thus far have dealt with modules of genes associated with either phenotypic or genotypic information. While both approaches are helpful for predicting dys-regulated modules, a

more effective way to understand disease mechanisms might be to combine both genotypic (the putative causes of diseases) and phenotypic data (their effects). Expression Quantitative trait loci (eQTL) analysis is a useful method to find the relationship between genotype and phenotype [72,73]. eQTL treats the level of gene expression as a quantitative phenotype, which is assumed to be controlled by genotypic information. Loci that putatively control the expression of a given gene are identified by determining the associations between genotype and gene expression. Given an association between a genotypic variation in a locus and expression level of a gene, the next challenge is to uncover the pathway(s) through which the genetic variation leads to the expression change. Recently, several groundbreaking pathway elucidation methods have emerged, as illustrated in Figure 3 and described below.

3.3.1 Distance based methods. A simple approach to identify a possible pathway from a genetically altered gene (putative cause) to the gene with correlated expression change (target gene) is to test if there is a path in an interaction network connecting the putative causal gene to its target gene. The shortest path connecting a causal gene and its target is often used to explain their causal relationship (Figure 3A). The intermediate nodes on such a shortest path are likely members of an affected pathway/module. Several variations of the shortest path approach have been used in extracting disease associated network modules [74–76]. For example, Carter *et al.* searched for the shortest directionally consistent paths in molecular interaction networks connecting

seed genes to their targets. The targets were inferred by linear decomposition of gene expression data [76].

When multiple target genes exist, the well-known graph-theoretical concept of a Steiner tree is often used in place of a set of shortest paths. Given a set of nodes to be connected, a Steiner tree is an acyclic subgraph (a tree) connecting all these nodes while using the minimum number of edges. In a Steiner tree, the individual path from the putative causal gene (the root of the tree) to each of the target genes does not need to be the shortest, but the size (i.e., the number of edges) of the whole tree is minimized. The Steiner tree approach has been used to find new functional associations for proteins [77]. Tuncbag et al. extended the approach to the Steiner forest problem (allowing multiple trees), applying it to proteomic data from glioblastoma multiforme (GBM). In their study, each tree was rooted in a different cell surface receptor and represented independent signaling pathways originated from this receptor [78].

Distance based methods, such as the shortest path approach or the Steiner tree method, have several shortcomings. In particular, they ignore the fact that a pair of genes may have multiple paths connecting them in a network. In addition they use network topology without considering additional data (e.g. gene expression) and assume that the shortest pathways are the most informative or most likely used paths, which may not always be the case.

3.3.2 Flow-based methods. In the information flow approach, genotypic variations are considered the source of perturbation, while genes with phenotypic changes are considered the targets of a perturbation pathway. Instead of finding single paths connecting source and targets, flow-based methods compute the fraction of flow going through each intermediate node/edge. Fraction of flow indicates the probability of using the given path in information propagation (Figure 3B). In the case of current flow approach, the network is modeled to mimic the behavior of current in an electronic circuit, where each edge has an associated resistance. The current flow network provides an efficient framework equivalent to a random walk, which is also often used for modeling information flow in biological networks (see discussion below). An important advantage of network flow approaches their ability to incorporate additional data (such as gene expression, confidence level of interactions, and functional associations of genes) to the probabilistic network models. By incorporating such additional data, network flow approaches can more confidently suggest information propagation pathways.

The information flow of biological networks has been used to predict protein functions, to prioritize candidate disease genes, and to find network centralities [7,79–88]. The flow-based approach is particularly useful for augmenting network information for eQTL analysis. Specifically, it can be used to pinpoint likely causal genes in genomic eQTL regions and to uncover genes involved in the propagation of information signals from such causal genes to their target genes.

There are several mathematical formulations that can be used to capture information propagation. In addition to the aforementioned current flow, other approaches include random walk and network flow. While mathematically different, many information propagation methods share a number of similar assumptions such as flow conservation (Kirchhoff's law). In the random walk method, a number of random walkers repeatedly start from a node. The likelihood of associating a gene in the network to a disease is estimated by the number of random walkers arriving at the gene. Gene expression correlation provides one way to compute the weight of a gene in the network which, in turn, provides the transition probability of the random walker. The network flow methods are closely related to the current flow approach. Unlike current flow, however, the network flow model resembles water-finding paths through pipes. Capacities are associated with pipes (edges) providing constraints on how much flow can go through each pipe.

Examples. Tu et al. [79] used the random walk approach to infer causal genes and underlying causal paths over a molecular interaction network for yeast knock-out experimental data. Current flow is an equivalent form of random walk that can be used in a more computationally efficient way [89]. Using this knowledge, Suthram et al. [80] developed the eQED method, which integrates eQTL analysis with molecular interaction information modeled as a current flow network.

Kim et al. further extended the eQED idea to identify causal genes and dysregulated pathways and applied it to Glioma sample analysis [69,90].One of the challenges of eQTL analysis is a massive multiple testing problem, for which various multiple testing correction methods have been proposed. Without such corrections, eQTL analysis typically finds multiple associated regions for each target gene, many of which are simply by chance. However, simply applying a more stringent p-value cutoff for multiple testing corrections often eliminates many true causal regions. Moreover, each region may contain dozens of candidate causal genes. Current flow analysis can be applied to complement eQTL analysis and help to identify the genes whose alterations are most likely to cause abnormal expression for the target gene. Using copy number variations and gene expression profiles of the same set of cancer patients, Kim et al. first identified chromosomal regions where copy number variations correlated with gene expression changes. Subsequently, they used the current flow algorithm to identify potential causal genes in the associated regions. By selecting genes receiving significant amounts of current in the network, Kim et al. identified putative causal genes in Glioblastoma and uncovered commonly dys-regulated pathways, including insulin receptor signaling pathways and RAS signaling. The identified pathways featured several hub nodes, such as EGFR, that were known to be important players in Glioma or more generally in cancer. Compared to simple genome-wide association studies, which only identify putative associations between causal loci and target genes, the current flow based method provides increased power to predict causal disease genes and to uncover dys-regulated pathways.

A variant of the network flow approach, the minimum cost network flow, was used to model the response to increased expression of alpha-synuclein, a protein implicated in several neurodegenerative disorders, including Parkinson's disease [81]. In addition to the edge capacities, the min cost network flow approach associates weights with edges representing the cost of sending flow through an edge. These weights were computed based on the probability of the two genes interacting in a response pathway, while capacities were calculated using the transcript levels of target genes.

4. Applications of Network Modules – Disease Diagnosis and Treatment

Can network modules help facilitate a more personalized approach for disease diagnosis and treatment? Traditional approaches of clinical disease classification have been based on pathological analysis of patients and existing knowledge of diseases. However, traditional diagnostic approaches are prone to errors. Alterna-

tively, knowledge about dys-regulated pathways can be used to subtype diseases and to develop relevant treatments for individual disease subgroups. For example, network modules have been used to predict patient survival, metastasis, drug responses for various types of cancer [61,68,91–94].

4.1 Disease Classification

A supervised approach to disease classification starts with a set of samples with a known partition into disease subtypes (e.g., metastatic or not) and attempts to identify a classifying principle using specific molecular features. The general strategy for supervised disease classification is to search for subnetworks, also called subnetwork markers, whose activities best discriminate the two disease subtypes. As in the case of single-gene disease markers, a network marker will distinguish some but not all disease cases and multiple subnetworks might be necessary. Among selected candidate network markers, the best markers are selected based on a set of training samples. Some methods take an unsupervised approach, where subclasses and their features are discovered without using a known training set.

Examples. Chuang *et al.* showed how dys-regulated network modules (described in Section 3.2) provide more robust and accurate predictions than those by single gene based classifications when applied to breast cancer metastasis analysis [61]. Chuang *et al.*'s work provided the proof of principle for using network modules in disease classification. A number of subsequent extensions and improvements to Chuang *et al.*'s work were suggested. For example, Lee *et al.* incorporated curated pathways, and searched for a subset of genes with discriminative features for the disease phenotype [94]. More recently, Dao *et al.* developed alternative network based approaches for classification of cancer subtypes by identifying densely connected subnetwork and randomized algorithms [92,93]. Other techniques for best marker identification, such as set cover and bottom-up enumeration techniques, were also proposed [68,91].

Kim *et al.* identified gene modules using a module cover approach to capture disease heterogeneity in brain cancer samples from Rembrandt and Ovarian Cancer samples from TCGA [70]. Next, Kim *et al.* superimposed the selected modules onto the results from an independently proposed classification scheme [57]. As a result, Kim *et al.* uncovered

which disease classes are characterized by which combinations of modules.

4.2 Disease Similarity

Network modules can also be used to explain disease similarity. Overlaps of dys-regulated network modules explain why some complex diseases share similar phenotypic traits. Suthram *et al.* used a variant of PathBlast [95] to identify dense subnetworks. Analysis of disease similarity was achieved by comparing expression patterns of various diseases in the modules [96]. Several dys-regulated modules were found to be common to many diseases, which explains why some drugs can treat many different diseases.

4.3 Response to Treatment

Modules may help determine whether a given patient will respond to a particular drug, which is valuable for treatment design. In addition, understanding molecular differences between responders and non-responders is likely to help development of alternative treatments. For example, Chu and Chen used a network approach to discover apoptosis drug targets [97]. Chu and Chen constructed a PPI network for apoptosis in normal cells and applied a nonlinear stochastic model to remove false positive interactions using microarray data. Comparing the resulting subnetworks helped to shed some light on the mechanisms leading to apoptosis and to identify potential drug targets.

5. Summary

Network biology provides powerful tools for the study of complex diseases. Network-based approaches leverage the idea that complex diseases can be better understood from the perspective of dys-regulated modules than at the individual gene level. Modularity is a widely accepted concept in molecular networks and module-based approaches provide a number of advantages including robustness in the identification of dys-regulated pathways and improved disease classification.

In addition, network based formulations allow using a wealth of methods already developed in graph theory, such as shortest paths, network flow, and Steiner trees. Network-based methods have several limitations including the lack of mechanistic explanations. Despite the limitations, network analysis has been applied successfully in many disease studies, suggesting testable hypotheses.

6. Exercises

1. Construct coexpression networks following the steps below [98].

a. Download the three expression datasets from the following page: http://www.geneticsofgeneexpression.org/network/download

b. Compute 3 population-specific correlations for each pair of 4238 genes with the expression data. (Hint: There are 8,978,203 pairs of genes.)

c. For gene pairs which have similar correlations in the 3 datasets, calculate the weighted average correlation, weighted by the number of individuals in each population. Hint: In the Supplemental Table 1 published with [98] (http://genome.cshlp.org/content/suppl/2009/10/02/gr.097600.109.DC1/nayak_supplemental_material.pdf), you can find the list of gene pairs whose correlations differ significantly among the 3 datasets.

d. Construct the correlation network by connecting gene pairs whose weighted average correlations are greater than a pre-defined threshold (e.g., 0.5).

e. Compute specific parameters describing the network topology. (Hint: You can use the NetworkAnalyzer Cytoscape plugin http://med.bioinf.mpi-inf.mpg.de/netanalyzer/)

f. For different correlation thresholds, compare the networks in terms of topological properties.

2. Suppose that in a co-expression network two genes are identified to have correlated expression patterns. Provide at least two possible biological explanations of this correlation.

3. Some variants of information flow approaches that identify pathways of information flow from a mutated gene to a target gene with correlated expression require that the last but one node gene on such a pathway (the node preceding the target gene) to be a transcription factor. What is a justification for such requirement? What can be advantages and disadvantages of such a design?

4. Consider a set cover approach to find a representative set of genes dys-regulated in a given set of cancer patients. The algorithm finds the smallest number of genes so that each disease case is covered at least k times. How does the number of selected genes depend on k? If you suspect that data for 5%
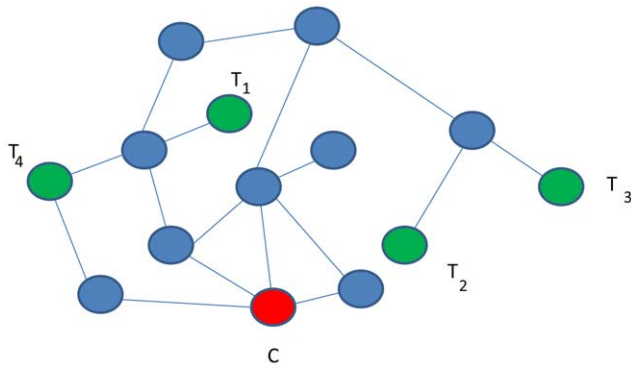
Figure 4. A hypothetical interaction network to be used with Exercises 5 and 6.
doi:10.1371/journal.pcbi.1002820.g004

patients might be incorrect, how would you modify the optimization problem?

5. A Steiner tree connecting a set of nodes does not need to be unique. In the graph shown in Figure 4, find two different Steiner trees connecting genes C, T_1, T_2, T_3, and T_4.

6. In the graph shown in Figure 4, find the shortest paths connecting C with each of T_1, T_2, T_3, and T_4. Do the edges used by these paths correspond to a Steiner tree? Explain why or why not.

Answers to the exercises are provided in Text S1.

Supporting Information

Text S1 Answers to Exercises.
(PDF)

Acknowledgments

Further Reading

- Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. Nature 461(7261): 218–223.
- Barabási AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12(1): 56–68.
- Przytycka TM, Singh M, Slonim DK (2010) Toward the dynamic interactome: it's about time. Brief Bioinform 11(1): 15–29.
- Przytycka TM, Cho DY (2012) Interactome. In: Meyers RA, editor. Encyclopedia of molecular cell biology and molecular medicine. John Wiley and Sons, Inc. doi:10.1002/3527600906.mcb.201100018
- Califano A, Butte AJ, Friend S, Ideker T, Schadt E (2012) Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet 44(8): 841–847. doi:10.1038/ng.2355
- Vidal M, Cusick ME, Barabási AL (2011) Interactome networks and human disease. Cell 144(6): 986–998.
- Kim Y, Przytycka TM (2012) Bridging the gap between genotype and phenotype via network approaches. Frontiers in Genetics special issue on mapping complex disease traits with global gene expression. Front Genet 3: 227. doi:10.3389/fgene.2012.00227

References

1. Veenstra-Vanderweele J, Christian SL, Cook EH, Jr. (2004) Autism as a paradigmatic complex genetic disorder. Annu Rev Genomics Hum Genet 5: 379–405.
2. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368–372.
3. Schadt EE (2009) Molecular networks as sensors and drivers of common human diseases. Nature 461: 218–223.
4. Gursoy A, Keskin O, Nussinov R (2008) Topological properties of protein interaction networks from a structural perspective. Biochem Soc Trans 36: 1398–1403.
5. Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118: 4947–4957.
6. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41–42.
7. Zotenko E, Mestre J, O'Leary DP, Przytycka TM (2008) Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. PLoS Comput Biol 4: e1000140. doi:10.1371/journal.pcbi.1000140
8. Jonsson PF, Bates PA (2006) Global topological features of cancer proteins in the human interactome. Bioinformatics 22: 2291–2297.
9. Wachi S, Yoneda K, Wu R (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 21: 4205–4208.
10. De Las Rivas J, Fontanillo C (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput Biol 6: e1000807. doi:10.1371/journal.pcbi.1000807
11. Berggard T, Linse S, James P (2007) Methods for the detection and analysis of protein-protein interactions. Proteomics 7: 2833–2842.
12. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, et al. (2008) High-quality binary protein interaction map of the yeast interactome network. Science 322: 104–110.
13. Shoemaker BA, Panchenko AR (2007) Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol 3: e43. doi:10.1371/journal.pcbi.0030043
14. Levy ED, Landry CR, Michnick SW (2009) How perfect can protein interactomes be? Sci Signal 2: pe11.
15. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
16. (2010) The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res 38: D331–335.
17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29.
18. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.
19. Costello JC, Dalkilic MM, Beason SM, Gehlhausen JR, Patwardhan R, et al. (2009) Gene networks in Drosophila melanogaster: integrating experimental data to predict gene function. Genome Biol 10: R97.
20. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, et al. (2008) A genomewide functional network for the laboratory mouse. PLoS Comput Biol 4: e1000165. doi:10.1371/journal.pcbi.1000165
21. Ramani AK, Li Z, Hart GT, Carlson MW, Boutz DR, et al. (2008) A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol Syst Biol 4: 180.
22. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7 Suppl 1: S7.
23. Peng J, Wang P, Zhou N, Zhu J (2009) Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc 104: 735–746.

24. Pe'er D (2005) Bayesian network analysis of signaling networks: a primer. Sci STKE 2005: pl4.

25. Alterovitz G, Liu J, Afkhami E, Ramoni MF (2007) Bayesian methods for proteomics. Proteomics 7: 2843–2855.

26. Xuan NV, Chetty M, Coppel R, Wangikar PP (2012) Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network. BMC Bioinformatics 13: 131.

27. Zou M, Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21: 71–79.

28. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47–52.

29. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22: 1021–1023.

30. Altaf-Ul-Amin M, Shinbo Y, Mihara K, Kurokawa K, Kanaya S (2006) Development and implementation of an algorithm for detection of protein complexes in large interaction networks. BMC Bioinformatics 7: 207.

31. Arnau V, Mars S, Marín I (2005) Iterative Cluster Analysis of Protein Interaction Data. Bioinformatics 21: 364–378.

32. Asthana S, King OD, Gibbons FD, Roth FP (2004) Predicting protein complex membership using probabilistic network reliability. Genome Res 14: 1170–1175.

33. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

34. Bader JS (2003) Greedily building protein networks with confidence. Bioinformatics 19: 1869–1874.

35. Brun C, Chevenet F, Martin D, Wojcik J, Guenoche A, et al. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biol 5: R6.

36. Dunn R, Dudbridge F, Sanderson CM (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. BMC Bioinformatics 6: 39.

37. Jiang P, Singh M (2010) SPICi: a fast clustering algorithm for large biological networks. Bioinformatics 26: 1105–1111.

38. King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20: 3013–3020.

39. Luo F, Yang Y, Chen CF, Chang R, Zhou J, et al. (2007) Modular organization of protein interaction networks. Bioinformatics 23: 207–214.

40. Navlakha S, Schatz MC, Kingsford C (2009) Revealing biological modules via graph summarization. J Comput Biol 16: 253–264.

41. Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci U S A 103: 8577–8582.

42. Pereira-Leal JB, Enright AJ, Ouzounis CA (2004) Detection of functional modules from protein interaction networks. Proteins 54: 49–57.

43. Qi Y, Balem F, Faloutsos C, Klein-Seetharaman J, Bar-Joseph Z (2008) Protein complex identification by supervised graph local clustering. Bioinformatics 24: i250–258.

44. Rives AW, Galitski T (2003) Modular organization of cellular networks. Proc Natl Acad Sci U S A 100: 1128–1133.

45. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci U S A 100: 12123–12128.

46. Wang C, Ding C, Yang Q, Holbrook SR (2007) Consistent dissection of the protein interaction network by combining global and local metrics. Genome Biol 8: R271.

47. Chen J, Yuan B (2006) Detecting functional modules in the yeast protein-protein interaction network. Bioinformatics 22: 2283–2290.

48. Feng J, Jiang R, Jiang T (2011) A max-flow based approach to the identification of protein complexes using protein interaction and microarray data. IEEE/ACM Trans Comput Biol Bioinform 8: 621–634.

49. Maraziotis IA, Dimitrakopoulou K, Bezerianos A (2007) Growing functional modules from a seed protein via integration of protein interaction and gene expression data. BMC Bioinformatics 8: 408.

50. Tipney H, Hunter L (2010) An introduction to effective use of enrichment analysis software. Hum Genomics 4: 202–206.

51. Kim Y, Przytycka T (2012) Bridging the gap between genotype and phenotype via network approaches. Frontiers in Genetics special issue on mapping complex disease traits with global gene expression. Front Genet 3: 227. doi:10.3389/fgene.2012.00227

52. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1: S233–240.

53. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485: 246–250.

54. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 18: 507–522.

55. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7: e1001273.

56. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, et al. Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70: 898–907.

57. The Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615.

58. Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, et al. (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70: 898–907.

59. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, et al. (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron 70: 886–897.

60. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7: e1001273. doi:10.1371/journal.pgen.1001273

61. Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3: 140.

62. Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. Nature 455: 401–405.

63. Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, et al. (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 4: 169.

64. Wang K NI, Banerjee N, Margolin AA, Califano A. Genome-wide Discovery of Modulators of Transcriptional Interactions in Human B Lymphocytes; 2006; Venice. pp. 348–362.

65. Xue H, Xian B, Dong D, Xia K, Zhu S, et al. (2007) A modular network model of aging. Mol Syst Biol 3: 147.

66. Xia K, Xue H, Dong D, Zhu S, Wang J, et al. (2006) Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. PLoS Comput Biol 2: e145. doi:10.1371/journal.pcbi.0020145

67. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R (2010) DEGAS: de novo discovery of dysregulated pathways in human diseases. PLoS ONE 5: e13367. doi:10.1371/journal.pone.0013367

68. Chowdhury SA, Koyuturk M (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. Pac Symp Biocomput: 133–144.

69. Kim YA, Wuchty S, Przytycka TM (2011) Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol 7: e1001095. doi:10.1371/journal.pcbi.1001095

70. Kim Y, Salari R, Wuchty S, Przytycka TM (2013) Module Cover – a new approach to genotype-phenotype studies; Pacyfic Synposium on Biocomputing 18: 103–110.

71. Vandin F, Upfal E, Raphael BJ (2012) De novo discovery of mutated driver pathways in cancer. Genome Res 22: 375–385.

72. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, et al. (2005) Genome-wide associations of gene expression variation in humans. PLoS Genet 1: e78. doi:10.1371/journal.pgen.0010078

73. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, et al. (2007) Population genomics of human gene expression. Nat Genet 39: 1217–1224.

74. Managbanag JR, Witten TM, Bonchev D, Fox LA, Tsuchiya M, et al. (2008) Shortest-path network analysis is a useful approach toward identifying genetic determinants of longevity. PLoS ONE 3: e3802. doi:10.1371/journal.pone.0003802

75. Shih YK, Parthasarathy S (2012) A single source k-shortest paths algorithm to infer regulatory pathways in a gene network. Bioinformatics 28: i49–58.

76. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, et al. (2007) Prediction of phenotype and gene expression for combinations of mutations. Mol Syst Biol 3: 96.

77. Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, et al. (2011) Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci U S A 108: 882–887.

78. Tuncbag N, McCallum S, Huang SS, Fraenkel E (2012) SteinerNet: a web server for integrating 'omic' data to discover hidden components of response pathways. Nucleic Acids Res 40: W505–509.

79. Tu Z, Wang L, Arbeitman MN, Chen T, Sun F (2006) An integrative approach for causal gene identification and gene regulatory pathway inference. Bioinformatics 22: e489–496.

80. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T (2008) eQED: an efficient method for interpreting eQTL associations using protein networks. Mol Syst Biol 4: 162.

81. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, et al. (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet 41: 316–323.

82. Lee E, Jung H, Radivojac P, Kim JW, Lee D (2009) Analysis of AML genes in dysregulated molecular networks. BMC Bioinformatics 10 Suppl 9: S2.

83. Kohler S, Bauer S, Horn D, Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 82: 949–958.

84. Missiuro PV, Liu K, Zou L, Ross BC, Zhao G, et al. (2009) Information flow analysis of interactome networks. PLoS Comput Biol 5: e1000350. doi:10.1371/journal.pcbi.1000350

85. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21 Suppl 1: i302–310.

86. Newman M (2005) A measure of betweenness centrality based on random walks. Social Networks 27: 39–54.

87. Stojmirovic A, Yu YK (2007) Information flow in interaction networks. J Comput Biol 14: 1115–1143.

88. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R (2010) Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 6: e1000641. doi:10.1371/journal.pcbi.1000641

89. Doyle PG SJ (1984) Random walks and electric networks

90. Kim YA, Przytycki JH, Wuchty S, Przytycka TM (2011) Modeling information flow in biological networks. Phys Biol 8: 035012.

91. Chowdhury SA, Nibbe RK, Chance MR, Koyuturk M (2011) Subnetwork state functions define dysregulated subnetworks in cancer. J Comput Biol 18: 263–281.

92. Dao P, Colak R, Salari R, Moser F, Davicioni E, et al. (2010) Inferring cancer subnetwork markers using density-constrained biclustering. Bioinformatics 26: i625–631.

93. Dao P, Wang K, Collins C, Ester M, Lapuk A, et al. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics 27: i205–213.

94. Lee E, Chuang HY, Kim JW, Ideker T, Lee D (2008) Inferring pathway activity toward precise disease classification. PLoS Comput Biol 4: e1000217. doi: 10.1371/journal.pcbi.1000217

95. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc Natl Acad Sci U S A 100: 11394–11399.

96. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, et al. (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. PLoS Comput Biol 6: e1000662. doi:10.1371/journal.pcbi.1000662

97. Chu LH, Chen BS (2008) Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. BMC Syst Biol 2: 56.

98. Nayak RR, Kearns M, Spielman RS, Cheung VG (2009) Coexpression network based on natural variation in human gene expression reveals gene interactions and functions. Genome Res 19: 1953–1962.