

19. Path Finding in Biological Networks

Lore Cloots, Dries De Maeyer, Kathleen Marchal

Understanding the cellular behavior from a systems perspective requires the identification of functional and physical interactions among diverse molecular entities in a cell (i.e., DNA/RNA, proteins, and metabolites). The most straightforward way to represent such datasets is by means of molecular networks of which nodes correspond to molecular entities and edges to the interactions amongst those entities. Nowadays with large amounts of interaction data being generated, genome-wide networks can be created for an increasing number of organisms. These networks can be exploited to study a molecular entity like a protein in a wider context than just in isolation and provide a way of representing our knowledge of the system as a whole. On the other hand, viewing a single entity or an experimental dataset in the light of an interaction network can reveal previous unknown insights in biological processes.

In this chapter we focus on different approaches that have been developed to reveal the functional state of a network, or to find an explanation for the observations in functional data through paths in the network. In addition we give an overview of the different omics datasets and

19.1	Background	289
19.2	Inferring Interaction Networks from Omics Data	290
19.2.1	Network Representations	290
19.2.2	Integration Frameworks Based on Supervised Techniques to Predict Interactions	292
19.3	Using Interaction Networks to Interpret Functional Data	294
19.3.1	Connecting One or Several Causes to Their Effect(s) by Unveiling the Underlying Active Paths	295
19.3.2	Integrating Cause–Effect Pairs to Confidently Infer Edge Attributes on the Network ...	300
19.3.3	Identifying (an) Unknown Causal Input(s) for (an) Observed Effect(s)	300
19.3.4	Identifying Network Entities Related to Network Entities of Interest	302
19.4	Conclusion	304
	References	304

data-integration techniques that can be used to build integrated biological networks.

19.1 Background

With the advent of new molecular profiling techniques, genome-wide datasets that describe interactions between molecular entities (i.e., mRNA, proteins, metabolites, etc.) are being generated at an ever increasing pace. These datasets each measure a specific type of interaction that is active under certain conditions within the cell or that occurs as a response to specific environmental signals. The distinct nature of these datasets often brings about complementary views on cellular behavior. A network-based representation of various biological systems captures many

of the essential characteristics of these data and integrating complementary molecular interaction layers into a single network thus provides a way of representing our knowledge of the system as a whole. Application of well-established tools and concepts developed in fields such as graph theory on such networks can provide valuable insights into the system's mode of action and functionalities [19.1]. The identification of motifs, which are statistically significant reoccurring characteristic patterns in a network [19.2], has for instance shown that specific types of motifs carry

out specific information-processing functions within cells [19.3].

An integrated interaction network can also reveal previous unknown insights in biological processes or functional behavior by explicitly interrogating it with independent functional data sets. Methodologies that identify and explore paths in networks between given input and output nodes have gained much interest. Such a path in a network can be seen as a mechanistic representation of the way information propagates through the network. Identifying biologically meaningful paths in the network between nodes of interest,

nodes which can be defined from functional data sets that are independent from the network itself, can unveil previously uncovered signal flow mechanisms that are responsible for the observed functional behavior or define a measure for relatedness of two nodes in the network.

In this chapter, we highlight diverse omics datasets and data-integration techniques that can be used to build integrated biological networks and discuss several categories of network-based path finding methodologies that aim at obtaining a more functional understanding of cellular behavior.

19.2 Inferring Interaction Networks from Omics Data

Understanding the cellular behavior from a systems perspective requires the identification of functional and physical interactions among diverse molecular entities in a cell (i.e., DNA/RNA, proteins and metabolites). The most straightforward manner of capturing interactions between molecular entities is by representing them as an interaction network. Here, molecular entities are represented by nodes and the interactions between them by edges. In this section, we elaborate on different types of networks that can be constructed from omics data and present supervised learning strategies to assign reliabilities to interactions.

19.2.1 Network Representations

Classically, a distinction is made between a functional network in which nodes usually correspond to proteins or genes, and edges represent functional relations between the nodes and a physical network where edges represent direct physical interactions (Fig. 19.1). Proteins connected in a functional network can be interpreted as being active in the same pathway or being needed together to mediate a specific function, but they do not necessarily physically interact. Examples of specific functional networks are, for instance, genetic interaction networks and coexpression networks. Within a physical network, different molecular layers can be distinguished: intracellular signal transduction, which transmits information from the surface to the nucleus, for instance by means of protein phosphorylation, and protein interactions that propagate this signal. In addition, (post-) transcriptional regulation processes comprise transcription factor (TF) proteins or sRNAs regulating the expression of genes and finally metabolic

reactions catalyzed by enzymes, where metabolites are converted into energy and building blocks. Each of these different layers in a physical network can be deduced from their own specific datasets and will have their own characteristics.

Overview of Different Datasets

Small-scale laboratory experiments alone are impractical for creating a genome-scale network of different types of interactions, mainly for reasons of cost and time. Recently, advances in experimental methods made it possible to generate interaction datasets in a high throughput manner. Such datasets, like for instance, protein-protein interactions (PPI), have been generated for model organisms such as *Saccharomyces cerevisiae* [19.4–8], *Caenorhabditis elegans* [19.9], and *Drosophila melanogaster* [19.10, 11], as well as *Homo sapiens* [19.12, 13] by genome-wide yeast two hybrid (Y2H) screens and large-scale affinity purification/mass spectrometry. Technologies such as ChIP-chip and ChIP-seq, make it possible to measure the TF-DNA interactions at a genomic scale [19.14–16], and mass spectrometry (MS)-based proteomics have enabled the large-scale mapping of in vivo phosphorylation sites [19.17].

Due to the large flood of experimental interaction data becoming available, several efforts have been made to store and centralize these datasets through the construction of databases. Some databases capture data about a specific organism or research topic, like, for instance, transcriptional regulation, while others integrate data from specific organisms and/or different interaction types in a standardized manner. Gradually more and more specific databases are merged into these integrated

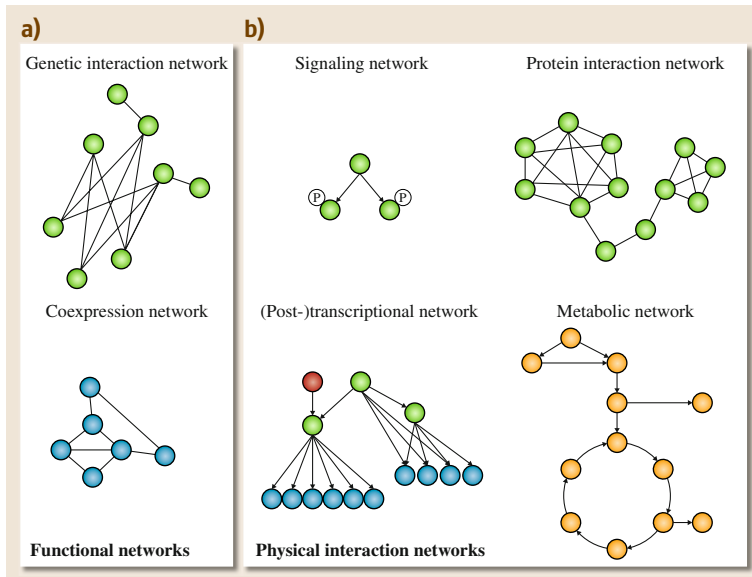


Fig. 19.1a,b Overview of molecular networks that can be inferred from omics data. **(a)** Functional networks: molecular entities represented by nodes in the network share a functional relation that does not require their physical contact. Genetic interaction network: edges reflect the phenotype that is observed when both nodes (genes) are inactivated simultaneously (double mutant). *Coexpression network*: nodes correspond to genes and edges represent the mutual similarity in expression profiles between connected nodes. **(b)** Physical interaction networks. Physical contact occurs amongst members (nodes) of the network. *Signaling network*. Nodes are proteins and edges represent signaling events (e.g., phosphorylation). *Protein interaction network*. Edges represent physical interactions between proteins represented by the nodes. *(Post-)transcriptional network*. Nodes represent either regulators or target genes and the directed edges reflect the physical regulator (transcription factor, sRNA)-target interactions. *Metabolic network*. Edges correspond to metabolic reactions catalyzed by enzymes represented by the nodes

databases. Table 19.1 gives an overview of some frequently used databases that provide physical interaction data for several eukaryotic model organisms, categorized by the type(s) of data they provide.

The fragmentation of data over a rising amount of databases [19.33] makes an integrated and comprehensive use very difficult. To reduce this problem, Bader, Cary, and Sander [19.34] provide an extensive overview of several databases (at the time of this writ-

ing this number was equal to 328) spanning different interaction layers across a multitude of organisms in a meta-database named Pathguide.

Assessing the Confidence in Interactions Derived from High-Throughput Experimental Data

Experimental datasets generated by high-throughput methods like Y2H or ChIP-chip are not only prone to

Table 19.1 Overview of databases containing physical interaction data for eukaryotic model organisms

Type of interactions	Database
Metabolic pathways	KEGG [19.18], MetaCyc [19.19], BiGG [19.20]
Protein-protein interactions	BioGRID [19.21], BIND [19.22], DIP [19.23], MINT [19.24], MIPS [19.25], STRING [19.26], HPRD [19.27] (<i>H. sapiens</i>)
Protein complexes	MIPS [19.25], CORUM [19.28] (mammalian organisms)
Transcription factor-DNA interactions	TRANSFAC [19.29], modENCODE [19.16] (<i>D. melanogaster</i> , <i>C. elegans</i>), YPD [19.30] (<i>S. cerevisiae</i>)
Signaling interactions	PhosphoPOINT [19.31], PhosphoSite [19.32]

high rates of false positive interactions [19.15, 35], low overlap between them also indicates that the current interaction maps are far from complete [19.36]. Assessing the quality of the data obtained is useful both for deciphering the correct molecular mechanisms underlying given biological functions, and for intelligent future experiment design [19.37].

Von Mering et al. [19.38] addressed the problem of extracting highly confident interactions between proteins from high throughput data sources by using the intersection of direct high throughput experimental results. Although they were able to achieve low false positive rates, the coverage in the number of retrieved interactions was also very low. Increasing the coverage of a network is especially useful for humans since the protein interaction map, for instance, was estimated to be only 10% complete [19.36]. Predicting interactions augments the current knowledge of the relationship between distinct cellular processes and underlying mechanisms of diseases.

Several machine-learning methodologies were suggested to assign reliabilities to interactions identified by experimental data, as well as to predict *de novo* interactions. In Sect. 19.2.2, we, therefore, introduce the concept of supervised learning to assess interactions.

19.2.2 Integration Frameworks Based on Supervised Techniques to Predict Interactions

Supervised learning methods [19.39] infer a function from a training set. Such a training set consists of pairs of input vectors and their corresponding known output. When the output is discrete, the supervised method is called a classifier. The learned function between input and output can then be used to predict the output of any valid input of which the output is not yet known. Applied here, a classifier would exploit known interactions to infer novel interactions from omics data. They can learn the set of data characteristics (features) that allow distinguishing true from false interactions from a set of known interactions (training set). A novel interaction is then predicted to be true or false, depending on the extent to which it shares similar features with the interactions in the training set.

Since classifiers are commonly used to assess the confidence in interactions from omics data, some guidelines for the choice of features and training sets are discussed in the next section, as well as two types of classifiers that are frequently used to stratify many candidate interactions by confidence or predict novel

interactions, namely Bayesian approaches and logistic regression. A short case study in predicting PPI and functional interactions is presented in the last part of this section, as an enormous amount of high throughput experimental data for PPI is nowadays freely available in several databases. However, the integration process is general and can be used for assessing interactions at other network levels as well, using the standard frameworks described below, together with a set of features and training set that is specific for the dataset being assessed (e.g., [19.35, 40] for assessing TF-DNA interactions obtained from CHIP-chip).

Features

The set of features provided to the classifier are measurable entities or *evidences* that characterize an interaction or a noninteraction. The classifier then learns which of the provided features are predictive for the interactions at hand. Such measurable entities can be direct information (e.g., the interaction was seen in an experiment) or indirect information (e.g., the expression correlation of two proteins could indicate that they are members of the same complex). Examples for predicting PPI are, for instance, network topology-based features [19.41] and GO biological process similarity [19.42, 43], amongst others. Nucleosome occupancy [19.40], DNA binding motifs [19.35, 40], and shared phylogenetic profiles (i.e., occurrence of the interaction in multiple species) [19.35] have been shown to be predictive for TF-DNA interactions. Coexpression between genes [19.35, 40, 42–45] has been used both for predicting TF-DNA and PPIs.

Training Sets

The prediction quality of a classification scheme stands or falls with the choice of a golden standard training set. This training dataset usually consists of positive and negative examples and is used to discover a predictive relationship between several features and the positive and negative examples.

An ideal golden standard should be independent of the data sources serving as features, sufficiently large for reliable statistics, and free of systematic bias [19.42]. Moreover, the choice of training set also depends on the prediction task at hand: positive and negative examples should reflect the same entities as the ones one would like to predict. This means, for instance, that a golden standard for predicting protein complexes should consist of proteins belonging or not belonging to the same complex, while positive and negative examples for predicting a functional network, on the other

hand, should reflect functional and nonfunctional relationships, respectively.

A set of positive examples is usually based on a curated, literature-derived dataset, containing only high-confidence interactions. For predicting physical protein interactions, a high quality subset of the Database of Interaction Proteins (i. e., **DIP** [19.23]) discovered by small-scale experiments or data from individually performed experiments listed in The Munich Information Center for Protein Sequences (i. e., **MIPS** [19.25]) can, for instance, be applied. **TF–DNA** interactions from the Incyte YPD database [19.30] could serve as a positive set for transcriptional interactions in yeast [19.35, 40]. Positive examples of functional relations between proteins can be extracted from the gene ontology (i. e., **GO** [19.46]) database annotations. Usually, proteins are considered functionally related if they share a specific biological process **GO** term (e. g., contain less than 200 annotations [19.47]).

A good set of negative examples is harder to define, since noninteracting pairs cannot be observed. Negative training sets can, for instance, consist of randomly combined pairs [19.35, 44], randomly observed interactions in a high throughput dataset [19.45, 48], or, in the case of protein interactions, proteins occurring in different subcellular components [19.42–44], and proteins not sharing any specific **GO** term [19.47, 49–53].

Bayesian Approaches

Different sources of evidence can be probabilistically combined to predict interactions using Bayesian formalism. This learning framework allows for combining highly dissimilar types of data in a model that is easy to interpret and that can readily accommodate missing data.

The posterior odds of interaction between two molecular entities (O_{post}) represent the probability that an interaction occurs given the presence of several genomic features, divided by the probability that such an interaction will not occur given the presence of these features. This can be formalized using Bayes' theorem,

$$O_{\text{post}} = \frac{P(I|f_1, \dots, f_N)}{P(\sim I|f_1, \dots, f_N)}, \quad (19.1)$$

$$= \frac{\frac{P(f_1, \dots, f_N|I) \cdot P(I)}{P(f_1, \dots, f_N)}}{\frac{P(f_1, \dots, f_N|\sim I) \cdot P(\sim I)}{P(f_1, \dots, f_N)}}, \quad (19.2)$$

$$= \frac{P(I)}{P(\sim I)} \cdot \frac{P(f_1, \dots, f_N|I)}{P(f_1, \dots, f_N|\sim I)}, \quad (19.3)$$

$$= O_{\text{prior}} \cdot \text{LR}. \quad (19.4)$$

The posterior odds of an interaction can thus be calculated as the product of the prior odds (O_{prior}) of interaction and the likelihood ratio (**LR**) of an interaction (19.4).

The prior odds of interaction are defined as the probability of encountering an interaction among all pairs, divided by the probability of observing no interaction between a pair. The likelihood ratio represents the probability of observing the values in the predictive datasets given that a pair of molecular entities interacts, divided by the probability of observing these values given that the pair does not interact.

A naive Bayesian classifier makes the assumption that the genomic features (denoted by f_1, \dots, f_N) are independent. In this case, the **LR** can be calculated as the product of the individual likelihood ratios from the respective genomic features (19.5)

$$\text{LR} = \prod_{i=1}^N \left[\frac{P(f_i|I)}{P(f_i|\sim I)} \right]. \quad (19.5)$$

The likelihood ratio for every genomic feature can be estimated by counting the frequency of occurrence of interacting and noninteracting pairs in the golden standard that possess a particular value of the feature.

In the case of features with correlated evidence, the likelihood ratio cannot be factorized in this way and all possible combinations of all states of the features must be considered, which can be computationally intensive. The prior odds are more difficult to assess, since not all true interactions are known. For **PPI** this parameter, for instance, be estimated by examining the average number of interactions per protein for which all known interactions have been identified in the literature [19.43, 44].

After deriving likelihood ratios for independent features from the golden standard, the likelihood ratio for every protein pair can be determined by combining the likelihood ratios for every independent evidence source [19.43, 44]. An interacting pair is then predicted as positive if its likelihood ratio exceeds a certain cut off [19.42, 49, 51].

Logistic Regression

A logistic regression is a generalized linear model that is used to calculate the probability of the outcome of an event, e. g., the probability of observing an interaction between two proteins. The relationship between the response variable (e. g., observing an interaction ($= 1$) or not ($= 0$)) and the predictor variables (i. e., genomic features and/or experimental observations) is given by

a logistic function

$$P(I) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 f_1 + \dots + \beta_N f_N)}} \quad (19.6)$$

The logistic function can take as an input (i. e., evidence features f_1, \dots, f_N) any value from negative infinity to positive infinity, whereas the output (i. e., probability of an interaction $P(I)$) is confined to values between 0 and 1. Parameters β_1, \dots, β_N can be estimated by using a set of positive and negative interaction pairs as output (yielding output values of 1 and 0, respectively) and their corresponding features as input in a maximum likelihood approach [19.39]. The estimated parameters can then be used together with evidence features corresponding to an interaction between two molecular entities to predict the probability that these entities truly interact.

Case Study: Inferring Protein Interaction Networks from Omics Data

Examples of PPI and Functional Networks. Using the Bayesian framework or slight variations of it and specific sets of genomic features or experimental datasets, both functional networks for *S. cerevisiae* [19.49, 50], *C. elegans* [19.54], human [19.53], *Arabidopsis thaliana* [19.52], mouse [19.47] and protein-complex [19.42], and PPI networks [19.43, 44] for human were developed.

Methods based on logistic regression have been used to assess the confidence of interactions observed

in experimental data [19.41, 45, 48] by integrating experimental information, topological measures and/or expression correlation, and have been used in several path finding approaches to assign a confidence score to the PPI in a yeast network [19.55–58] or a human network [19.59], and to assess the reliability of experimentally determined protein interactions in *D. melanogaster* [19.11].

Performance. Many supervised classification methods have been developed to integrate direct and indirect information on protein interactions. They each differ in the collection of integrated data sources, approach, and implementation. Qi et al. [19.60] independently investigated the performance of different classifiers and the importance of different biological datasets, together with various golden standards. They concluded that a classifier based on Random Forests performed best among the classifiers, followed by a logistic regression. However, Suthram et al. [19.61] assessed the performance of six approaches, each with their own combination of features and classification method, and showed that a rather complex approach based on Random Forests [19.62] had lower overall performance compared to other methods tested. Both authors could conclude that including many input variables does not necessarily result in a better prediction performance, and in some cases even the opposite can be true. However, utilizing any probability scheme turned out to be better than considering all interactions observed to be true or equally probable.

19.3 Using Interaction Networks to Interpret Functional Data

Nowadays with large amounts of interaction data being generated, genome-wide networks can be created for an increasing number of organisms. These networks can be exploited to study a molecular entity like a protein in a wider context than just in isolation. However, the inferred physical networks are static and do not reveal which parts of the networks are active under certain conditions and how perturbations are propagated through the network. Integrating physical interaction networks with functional data, like gene expression data makes it possible to reveal relevant active paths or substructures in the network.

High-throughput techniques now allow genome-wide views of the molecular changes that occur in cells as they respond to stimuli. However, data derived

from these high-throughput techniques unveil that our understanding of cellular systems is still fragmentary, even of well-characterized model systems. In humans, for instance, only about 30–40% of all differentially expressed genes for transcription factors NF- κ B and STAT1 appear to be direct targets [19.63]. Yeager–Lotem et al. [19.64] observed that the results of genetic screenings (i. e., identifying genetic hits, or genes whose individual manipulation alters the phenotype of stimulated cells) and mRNA profiling (i. e., identifying differentially expressed genes following stimuli) often hardly overlap and provide a limited and biased view of cellular responses.

Exploiting the network structure can help in gaining a comprehensive picture of the functioning of a cell,

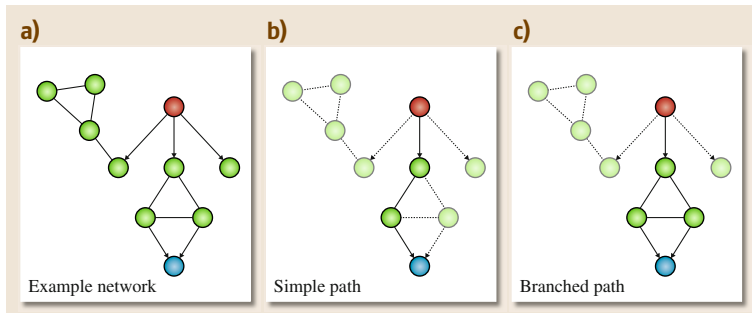


Fig. 19.2a–c Definition of a path (a) example of an interaction network, with an input node (*red node*), output node (*blue node*), and other interacting entities (*green nodes*) mapped on the network. *Arrows* between nodes represent directed interactions, *lines* between nodes represent undirected interactions (b) a simple path from input to output node is highlighted in the network, *dashed lines*, *arrows*, and *transparent nodes* do not belong to the selected path (c) a branched path from input to output node, containing multiple simple paths is highlighted in the network. *Dashed lines*, *arrows*, and *transparent nodes* do not belong to the path

by providing a mechanistic explanation that links the observed effects to the perturbation or cause exerted. Such an underlying mechanism is unlikely to be discoverable when looking at all datasets separately. Several approaches exist for mining the information embedded in integrated networks, dependent on the specifics of the problem statement. Network clustering strategies for instance, which search for highly connected subnetworks, have been successfully used to distinguish cancer-causing mutations from neutral mutations [19.65] or to assess the structure of the yeast genetic interaction network, revealing insights in gene function and modular organization [19.66].

In this chapter we focus on different approaches that have recently been developed to reveal the functional state of a network or to find an explanation for the observations in functional data through paths in the network. A simple path in a network is illustrated in Fig. 19.2b and is defined as a collection of edges that connect a source node (i. e., gene causing an effect or input gene of interest) and a target node (i. e., affected gene or output gene of interest) in an interaction network, such that each selected edge is connected to one other selected edge and the information spread by the source node can reach the target node without interruption. A path can be a collection of simple paths, containing several branches connecting a source with a target node (Fig. 19.2c). There is no further constraint that the nodes within a path should be densely connected to each other, which would refer to a cluster in a graph and would comprise a different problem statement.

In this second part of the chapter, an extensive overview of path finding methodologies, illustrated with

several applications, is given. Different approaches are categorized according to the underlying goal they try to accomplish. These goals are represented in an abstract way in Figs. 19.3–19.6, and are further clarified at the beginning of each category.

19.3.1 Connecting One or Several Causes to Their Effect(s) by Unveiling the Underlying Active Paths

The common objective of methods described in this paragraph is to reveal the underlying pathways transmitting a signal from one or several causes to their corresponding observed effect(s) by adopting a network-based approach (Fig. 19.3). The cause could, for instance, be a membrane protein and the observed effect a DNA binding protein that receives the signal, but the intermediate molecular interactions through which the signal was transduced from cause to effect is unknown.

Several of the methods developed for this purpose use, in addition to the given cause and effect pairs, other functional data like gene expression to extract biologically relevant paths from the network. This either by using the extent to which a network node is differentially expressed as an indication of its contribution to a plausible signaling path [19.64, 67] or by using a measure of expression correlation between edge nodes [19.55, 68], between edge nodes and the source and target nodes [19.69], or between edge nodes and target node [19.70] to indicate the confidence we have in an edge contributing to a causal path.

The reconstruction of signaling pathways by overlaying PPI data with cause–effect pairs has received

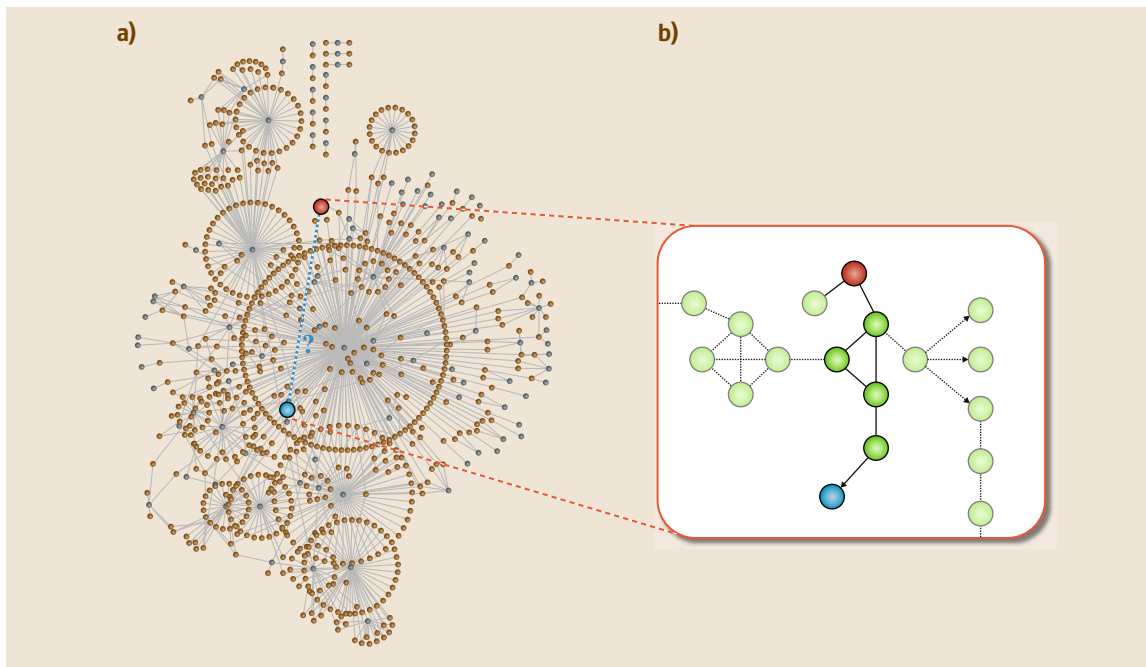


Fig. 19.3a,b Connecting one or several causes to their effect(s) by unveiling the underlying active paths. **(a)** Example of an interaction network on which a known causal gene (*red node* or input) and its affected gene (*blue node* or output) were mapped. The underlying path responsible for transferring the information from input to output is unknown (*blue dashed line*). **(b)** The underlying mechanism that explains the observed effect is highlighted in the network, *dashed lines, arrows, and transparent nodes* do not belong to the selected path. *Arrows* between nodes represent directed interactions, *lines* between nodes represent undirected interactions

a great deal of attention. Steffen et al. [19.71] were one of the first to model simple paths of a specified length through a physical protein interaction network, starting at a membrane protein and ending on a DNA binding protein in a procedure called NetSearch. Paths were ranked based on a statistical scoring metric, reflecting how many path members clustered together according to their expression profiles. Simple paths that had common starting points and endpoints and the highest ranks among each other were then combined into the final model of branched networks.

In reality, simple paths cannot capture the full complexity of signaling pathways since there may be multiple interaction paths within a pathway. Scott et al. [19.55], therefore, adapted the color coding technique and allowed the identification of more complicated substructures such as trees and series-parallel graphs. A number of candidate paths are firstly found with a score assigned to each candidate and the top scoring paths are then assembled into a signaling network. Lu et al. [19.69] extracted nonlinear path structures

from the network and potential interactions between related paths were taken into account.

However, most of these methods generally cannot directly find a signaling network as a whole, i. e., they first identify separate paths and then heuristically assemble them into a signaling network. Other approaches like those of Zhao et al. [19.68], Yosef et al. [19.58], Yeger-Lotem et al. [19.64], and Ren et al. [19.70] infer active paths immediately as a subnetwork from the whole network. The methods have in common that they try to explain cause–effect pairs in a particular set of experiments by solving an optimization problem which typically balances the reliability of the edges used by the length and complexity of the possible paths. A third category of methods uses the frequency of occurrence of a path with a predefined form (i. e., a motif) in the network to explain cause–effect pairs on a more statistical basis. An example of this category is the method of Joshi et al. [19.72], which is discussed in more detail in the case study at the end of this section.

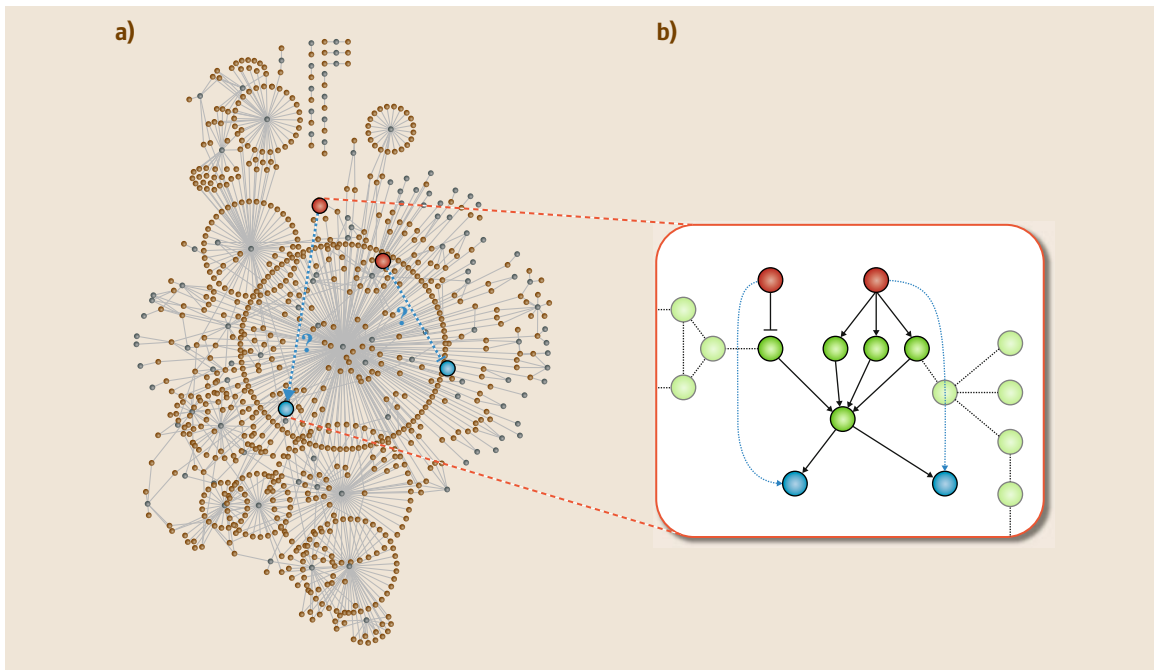


Fig. 19.4a,b Integrating cause–effect pairs to confidently infer edge attributes on the network **(a)** example of an interaction network on which different known cause (*red nodes*)–effect (*blue nodes*) pairs were mapped. The type of observed effect is also taken into account: a *regular arrow* represents an activating effect from input to output; a *cut arrow* represents an inhibiting effect from input to output. The underlying path responsible for transferring the information from input to output is unknown (*blue dashed line*). Also the type of effect (i. e., activating or inhibiting) for each edge on the path must be inferred consistently by making use of the cause–effect pairs **(b)** the underlying path that explains the observed effect is highlighted in the network, whereby also a type of effect to each edge in the path is assigned. *Dashed lines, arrows, and transparent nodes* do not belong to the selected path, the *blue dashed lines* show the observed cumulative effect from input to output as in **(a)**. *Arrows* between nodes represent directed interactions, *lines* between nodes represent undirected interactions

The majority of these methods use one or more MAP kinase signaling pathways involved in pheromone response, filamentous growth, maintenance of cell wall integrity, and high osmolarity as their benchmark, since these are among the best studied signaling networks. These pathways are activated by G protein-coupled receptors and characterized by a core cascade of MAP kinases that activate each other through sequential binding and phosphorylation reactions. A method comparison performed by Zhao et al. [19.68] demonstrated that most methods can to a large extent uncover the known signaling paths, which confirm the effectiveness and prediction power of the approaches. On the other hand, the results also show that there is no single method that can perform the best in all cases, and different models are complementary to each other.

While previous methods concentrate on the reconstruction of signaling cascades between a membrane receptor protein and a target protein, Yeager–Lotem et al. [19.64] focus on identifying molecular interaction paths connecting several related genetic hits (sources) and differentially expressed genes (targets), revealing the underlying response pathways. They hypothesize that some of the genetic hits, which are enriched for regulators of cellular response, will be connected via regulatory paths to the differentially expressed genes, which are the outputs of such paths, via components of the response that were not detected by either the genetic or the mRNA profiling assays themselves. To identify these undetected path components, the authors developed a flow algorithm called ResponseNet. Huang and Fraenkel [19.67] reconsidered this problem by taking into account that both the input data from experimen-

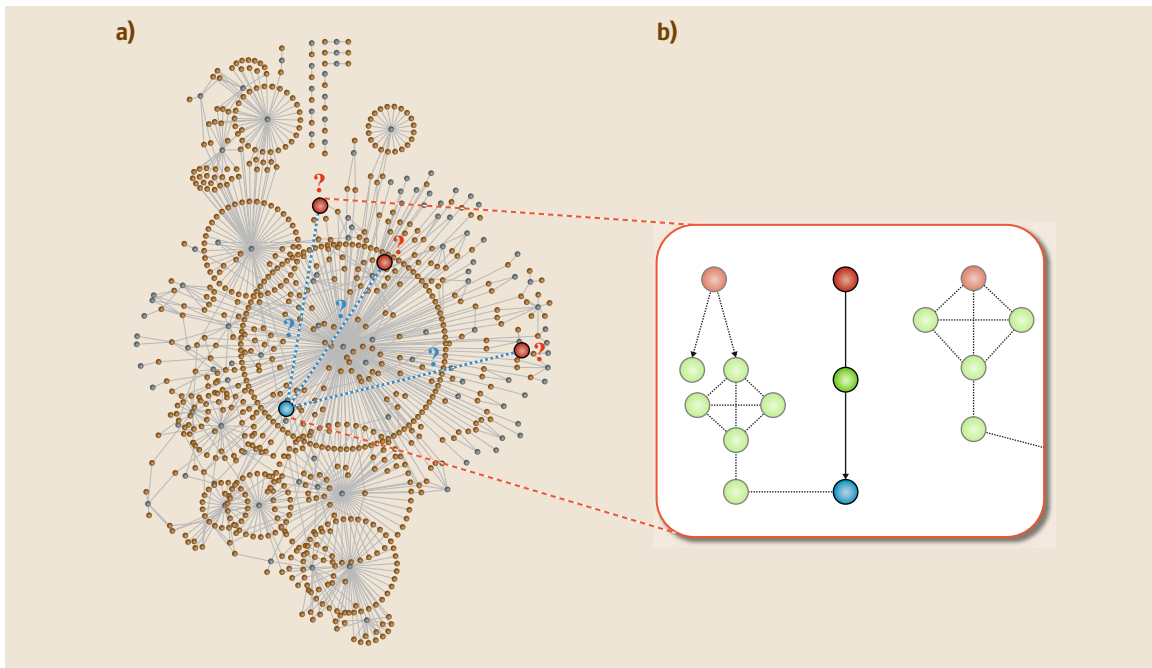


Fig. 19.5a,b Identifying (an) unknown causal input(s) for (an) observed effect(s). **(a)** Example of an interaction network on which several candidate causal genes (*red nodes* or possible inputs) and an affected gene (*blue node* or output) were mapped. The underlying path responsible for transferring the information from input to output is unknown (*blue dashed line*). Also the most likely input for the observed output should be identified (*red question marks*). **(b)** the most likely causal gene (*red node*) together with the underlying mechanism that explains the observed effect (*blue node*) is highlighted in the network, *dashed lines* and *arrows* and *transparent nodes* do not belong to the selected path. *Arrows* between nodes represent directed interactions, *lines* between nodes represent undirected interactions

tal observations and the interactome can contain noise. They treat the goal of connecting data as a constraint that is attempted to be satisfied through an optimization procedure, resulting in a subnetwork that contains mainly reliable edges while excluding possibly false positive source or target nodes.

Other application examples of this problem formulation can be found in the reconstruction of metabolic pathways [19.73], connecting a source metabolite to a target metabolite, and the reconstruction of transcriptional regulation [19.74], connecting regulators to their target module consisting of coexpressed genes.

Case Study

Previously mentioned techniques do not search for general mechanisms or path structures that are common between different cause and effect pairs, nor include a significance analysis that assesses the statistical significance of the inferred paths. In this way, it is difficult to assess if a given network model truly reflects under-

lying regulation mechanisms or appears just by chance due to the inevitable noise inherent in the perturbation data as well as in the physical interaction networks.

Joshi et al. [19.72], therefore, propose an alternative strategy by searching for *regulatory path motifs* in an integrated transcriptional (TRI), protein–protein (PPI) and phosphorylation (PhI) interaction network. Regulatory path motifs are defined as paths of length up to three, which connect a causative gene (for example a transcription factor) to a set of effect genes which are differentially expressed after perturbation of the causative gene, and occur significantly more often than expected by chance in an integrated physical network. The method was tested by searching regulatory motifs between 157 deleted [19.75] and 55 overexpressed [19.76] TFs in *S. cerevisiae*, together with their corresponding differentially expressed genes.

The significance of the regulatory path motifs is determined by a randomization strategy: the cause and

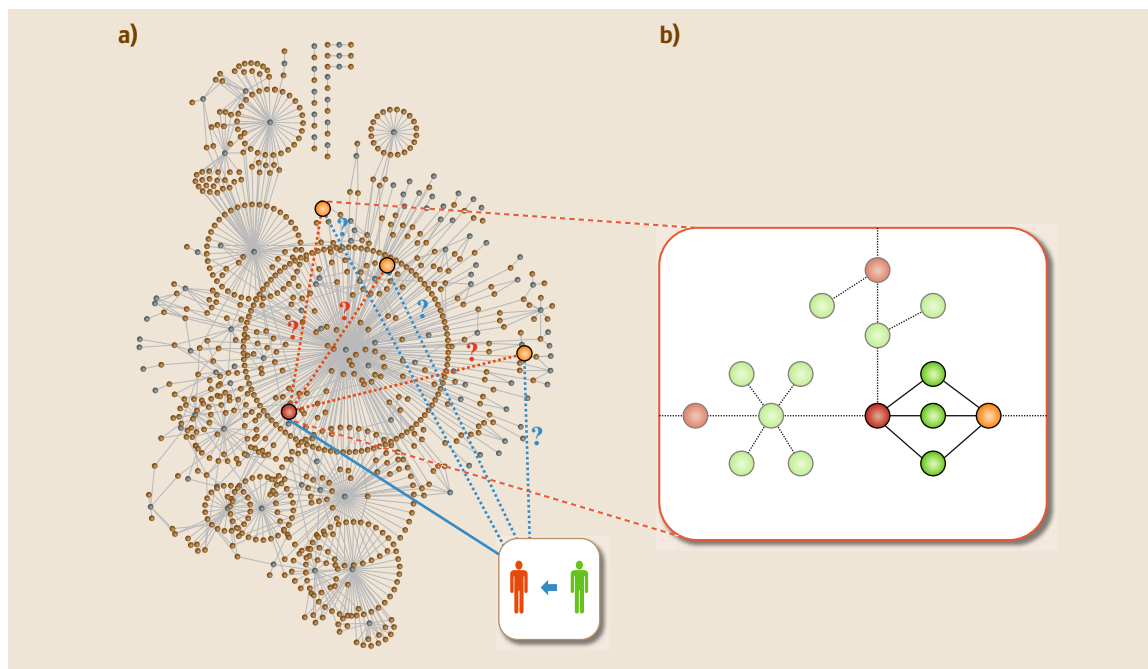


Fig. 19.6a,b Identifying network entities related to network entities of interest. **(a)** Example of an interaction network on which a known disease-related gene (*red node and full blue line to the disease*) and candidate disease-related genes (*orange nodes and dashed blue lines to the disease*) were mapped. To infer the most likely candidate disease-gene(s), their relatedness to the known disease-related gene (*dashed orange lines*) is examined through interactions on the network. **(b)** The most likely candidate disease-related gene and its relation to the known disease-related gene is highlighted on the network. *Dashed lines and arrows and transparent nodes do not belong to the selected paths*

effect pairs were permuted for 10 000 times, keeping the number of perturbed genes for each transcription factor constant. Next, the frequency of occurrence of each path motif in the randomized data sets was calculated. If the number of paths in the real perturbation data lies at the right tail of this random distribution (using a z-test statistic), the path was considered significant.

Out of all possible paths of length up to three, the algorithm identified eight regulatory path motifs, of which five were enriched in both deletion and overexpression data. These eight motifs explain 13% of all genes differentially expressed in deletion data and 24% in overexpression data, a more than five to tenfold increase compared to using directional transcriptional links only, confirming that perturbational microarray experiments contain mostly indirect regulatory links.

Like static network motifs [19.2, 77, 78], regulatory path motifs were found to aggregate into modular structures where the differentially expressed targets of a transcription factor reached by the same path through the same intermediate nodes, form a module. Many path

modules showed a high coexpression and were overrepresented in a particular functional category, validating the biological relevance of the regulatory path motifs.

This approach is not only more likely to reflect general regulatory strategies used in biological networks, but also the specificity of a TF to a particular regulatory path can hint towards its mode of action. It was found for instance, that 75% of the genes being perturbed after MET4 overexpression, can be explained by TRI, PPI-TRI, and PPI-TRI-TRI motifs, indicating that MET4 acts together with different combinations of auxiliary factors. In addition they observed that many network motifs that were significantly enriched in response to DNA damage in yeast were shorter than those enriched during cell cycle, exemplifying that environmental responses prefer fast signal propagation while developmental processes progress through multiple stages of interconnecting TFs [19.72, 79]. Thus regulatory path motifs can be used to characterize the condition-dependency of the response mechanisms across multiple integrated networks.

19.3.2 Integrating Cause–Effect Pairs to Confidently Infer Edge Attributes on the Network

While previous approaches limit themselves to inferring common underlying paths connecting related cause–effect pairs, others integrate several (even unrelated) known cause–effect pairs in order to assign specific attributes to the interaction graph (Fig. 19.4). Such attributes are, for instance:

1. The presence or absence of an edge in the path connecting cause and effect
2. The regulatory effect of a node, i. e., activating or repressing
3. The direction of information flow through an edge.

By integrating several cause–effect pairs more assignments of attributes can be made than when considering each pair in isolation. This because the assignment of attributes explaining a particular cause–effect pair should also be able to explain causes and observed effects that occur downstream of this pair in the network. Or in other words, the objective is trying to explain as many cause–effect pairs as possible such that the biological constraints on the network are consistent.

The inferred models of *Yeang et al.* [19.80], called physical network models, are annotated molecular interaction graphs. In this framework the presence of a certain edge in the physical network, the directionality of signal transduction in PPIs, and the regulatory effect of the interaction are determined if their combination is able to explain observed differential expression upon single gene knockouts. These strategies were further explored to investigate the mechanisms of the coupling between regulatory and metabolic networks [19.81]. The MTO algorithm of *Medvedovsky et al.* [19.82] and the method of *Gitter et al.* [19.83] limit themselves to determining a single direction for each edge, so that a maximum number of pairs have a directed path from the cause to the effect.

Case Study

SPINE [19.57] improved the physical network models [19.80] by assigning an activation/repression attribute with each protein so as to explain (in expectation) a maximum number of knockout effects observed. They do not explicitly model the direction of the edges, but most PPIs appeared in one direction only in the inferred consistent pathways. The goal of the algorithm is to infer regulatory pathways in the network that provide a consistent explanation for the input set of knockout pairs.

A path is a consistent explanatory path (Fig. 19.4b) if:

1. The aggregate sign of the path is equal to the observed expression direction (upregulated or activated versus downregulated or inhibited)
2. If every subpath connecting another knockout pair is also consistent.

The optimization problem is defined as that of finding an assignment that will maximize the expected number of pairs that have at least one consistent path, given by

$$E \left(\sum_{(s,t) \in X} K_{s,t} \right) = \sum_{(s,t) \in X} E(K_{s,t}) \\ = \sum_{(s,t) \in X} p(K_{s,t} = 1), \quad (19.7)$$

where $K_{s,t}$ is a variable that indicates if there exists at least one regulatory path consistent with a knockout pair (s, t) out of a collection of knockout pairs X ; $p(K_{s,t} = 1)$ corresponds to the probability that at least one consistent path exists for knockout pair (s, t) . The optimization problem is reformulated and solved as an integer linear program.

The authors evaluated their method by applying it on a genome-wide integrated yeast network consisting of PPI and TF-DNA data, in order to explain the effects observed in gene expression under different single-gene knockouts [19.84]. Here, a significant overlap between the model's prediction and the known signs was seen. Moreover, increasing the path length from one edge (i. e., only a direct TF-DNA link) up to three edges (i. e., one TF-DNA link and two other PPI/TF-DNA links) in different runs clearly showed the importance of looking at paths rather than considering direct edges only, since the amount of explained knockout pairs increased accordingly.

19.3.3 Identifying (an) Unknown Causal Input(s) for (an) Observed Effect(s)

In this section, we discuss applications where several effects are observed, but the true cause of these effects is unknown (Fig. 19.5). The common objective here is thus to infer this unknown cause or causes by use of a network. Examples of such problems can be found in the domain of expression quantitative trait loci (eQTL) mapping. With the availability of complete genomes of single strains, identifying which alterations in the DNA sequence (i. e., causes) are responsible for observed

changes in gene expression (i. e., the quantitative trait, or effects) becomes increasingly important. Usually, in a first step the association between a gene's expression level (i. e., **eQTL** mapping) and each genomic region (i. e., the expression quantitative locus or **eQTL**) is examined by a statistical method [19.85]. In the case of multifactorial traits, multiple loci can be associated with the gene's expression behavior, which complicates **eQTL** analysis. In addition, due to linkage disequilibrium each of the associated loci can contain several genes, which limits the localization of the true causal gene. Even when the causal gene can be identified, the molecular mechanism through which the association is exerted often remains elusive [19.86].

Path finding methods can here be used to identify the causal gene within an associated genomic locus and the underlying pathways that transmit signals from the locus to the affected target. The genes that altered their expression levels are considered as effects, while all the genes in the associated loci are defined as possible causes.

Tu et al. [19.87] proposed a random walk approach to infer the causal gene in a locus and the underlying pathways from a physical interaction network consisting of protein phosphorylation, **PPI**, and **TF-DNA** interactions. They assumed that the pathway starts with one causal gene in an associated locus and ends at the transcription factors regulating the target gene such that the expression of the genes on the pathway are correlated with the target gene. For each affected gene and each of its associated **eQTLs** their stochastic algorithm is performed separately to identify the causal gene. During the walks initiated on the network, different genes will be visited with different frequencies depending on their expression profile. The genes with higher frequencies are then assumed to be more likely to be the causal gene, and the most frequently traveled paths are regarded as the underlying regulatory pathways. For 239 out of 585 **eQTLs** identified in a study of 112 yeast segregants of Brem et al. [19.88] a causal gene could be significantly predicted. The authors highlighted GPA1 as causal regulator for target gene PRP39, a result that was experimentally verified by Yvert et al. [19.89].

Suthram et al. [19.86] further adapted the method of Tu by considering the analogy between random walks and electric circuits in a new method, named **eQTL** electrical diagrams (**eQED**). **eQED** models the flow of information from a locus to a target gene as electric currents through the protein network. The authors consider all loci influencing the target simultaneously, allowing multiple loci to reinforce each other when they fall

along a common regulatory pathway. The causal gene in each locus is then predicted as the one with the highest current running through it. By validating the **eQED** model on the **eQTL** data set of Brem and Kruglyak against a golden standard of knockout expression profiles, the multilocus model indeed showed a highly improved accuracy compared to the single-locus model and the method of Tu et al. (80% versus 72% and 50%, respectively).

Case Study

Inspired by the **eQED** electric circuit model, Kim et al. [19.90] developed a method for the identification of candidate causal genes and dysregulated pathways that are potentially responsible for the altered expression of target genes associated to glioblastoma multiforme (**GBM**), the most common and most aggressive malignant primary brain tumor in humans. Applied on gene expression and genomic alteration (in this case copy number variations or **CNVs**) profiles of 158 **GBM** patients, their methodology comprises four steps:

1. In a first step a set of genes is selected that show differential expression in the patients while taking into account disease heterogeneities among different patients, thus extracting sets of differentially expressed genes that are specific to a subgroup of patients. These differentially expressed genes are hereafter called target genes.
2. Subsequently, an **eQTL** mapping is performed by a linear regression analysis to determine the association between the expression of each target gene and copy number alterations of tag loci. A liberal *p*-value was chosen to retain most of potentially interesting relationships.
3. Then, to filter out false positives and to determine the most likely causal genes within each region of associated **CNVs** for each target gene, a physical network-based approach based on the electric circuit diagrams of Suthram et al. [19.86] was applied. Each node in the circuit represents a gene and holds a certain voltage to be determined. Each edge represents an interaction between node entities and has a conductance (i. e., how easily electricity flows along a path) defined by the mean expression correlation of its nodes with the target gene. As such, the authors ensured that a single noncorrelated node reduced but not completely interrupted the current flow, while a cluster of noncorrelated nodes put a considerable resistance to the current flow. Using Ohm's law and Kirchhoff's current law, the amount

of current through a node was calculated. Candidate causal genes for each target gene were then selected based on a permutation test to estimate the statistical significance of the current flow through the nodes.

4. Finally, this resulting set of causal genes was further reduced by imposing another filter: a minimum set of causal genes was selected that could explain all disease cases except for a few outliers.

Assessing the significance of the identified set of causal genes by determining overlap with sets of known GBM/glioma specific genes showed that their approach could uncover more cancer relevant genes than a simple association approach and demonstrated the increased predictive power of the model.

The authors also assessed the importance of genes in the paths from putative causal genes to their target genes and observed the emergence of hubs, genes that appeared in a disproportionately large number of paths. Such a set of hubs contained important transcription factors such as MYC and E2F1, and oncogenes such as JUN and RELA, and was enriched in genes that appeared in cancer pathways, the cell cycle, and several important signaling pathways. While such hub genes were clearly related to cancer, they would hardly have been identified by analyzing differentially expressed genes alone, demonstrating the advantages of a pathway-based approach.

Moreover, a GO biological process enrichment analysis of the uncovered subnetworks revealed frequently re-occurring classical cancer related pathways like insulin receptor signaling pathways, RAS signaling, as well as a glioma-associated regulation of transforming growth factor- β 2 production and SMAD pathway. Such pathways can then be considered as *GO biological process hubs* or *highways*, connecting many different causal genes with their targets. Such an observation supports the hypothesis of a pathway-centric view of complex disease, namely that many different genomic alterations potentially dysregulate the same pathways in complex diseases.

Among the discovered set of putative causal genes and pathways, an influence of PTEN and CDC2 was observed on the expression of WEE1 through transcription factors TP53 and E2F4. This tyrosine kinase in turn phosphorylates the protein product of CDC2 (i.e., CDK1), a signaling event that is crucial for the cyclin-dependent passage of various cell cycle checkpoints and suggested as an important feedback mechanism for cancer by the authors.

19.3.4 Identifying Network Entities Related to Network Entities of Interest

In this fourth and last category, the objective is to identify entities related to a set of entities of interest (hereafter named seeds), by exploring the paths (e.g., long versus short paths, paths through highly connected nodes versus through very specific nodes, one simple path versus multiple paths connecting cause and effect, ...) that connect them in the network, rather than inferring the underlying paths that transfer the signal from a cause to affected genes like in the previously described approaches (Fig. 19.6). This approach is useful when one is looking for the cause of an observed effect, but they cannot both be mapped on a physical or functional molecular network. A domain where this strategy, for instance, has been proven useful is when dealing with diseases as observed effects.

For most diseases only a limited number of causal genes is currently known [19.91]. Genome-wide association studies, whereby genomic variation are associated to a certain phenotype, typically result in one or more linked chromosomal regions, which in turn can contain several genes. Since the elucidation of disease mechanisms can improve diagnose or medical care, several approaches have been developed to identify novel disease genes.

Motivated by the observation that genes causing a specific or similar disease phenotype tend to lie close to each other in a protein-protein network [19.92], several network-based approaches have been developed. These methods have a common approach in the sense that they try to score candidate disease genes based on the assumption that good candidates reside in the neighborhood of certain a priori determined genes. These a priori determined genes are, for instance, genes known to be involved in the disease [19.93–95] or in related phenotypes [19.59, 96, 97], or differentially expressed genes upon the phenotype [19.98]. These a priori determined disease related genes are called *seed genes* in what follows.

Several types of measures can be applied to score candidate genes. A first, intuitive way of identifying disease related genes is based on direct neighborhood: candidate genes that are directly connected to one or more seed genes are then predicted to be potentially causative [19.93]. However, it is possible that two disease related genes do not interact with each other directly, but are, for instance, part of the same pathway, and disruption in either one of them leads to the same

disease. These cases will be missed by direct neighborhood counting. To account for indirect interactions, one can use the shortest path length between a seed node and a candidate gene as a measure of its relatedness to the seed gene: if a certain candidate node lies at most k edges away from the seed node, it is considered as a disease gene. *George et al.* [19.94] observed that as the shortest path length between a seed node and a candidate node increases, the sensitivity of identified disease genes improves, but the number of false positives increases exponentially and reduces the specificity.

Protein–protein networks, however, possess the *small world* property, meaning that the average path length between any two nodes in the network is rather short [19.99]. One consequence of this for methods relying on direct interactions or shortest paths is that it is not very unlikely to observe genes interacting with the disease seed genes but that are unrelated to the disease as such [19.95]. Moreover, methods based on shortest path length ignore the fact that there might be multiple shortest paths or also other paths with longer lengths, which could point to a higher relatedness to the seed gene than when only one path is present between seed and candidate node.

To overcome these limitations, several methods have been developed that consider the topology of the entire network (i. e., global distance measure versus local) to define a distance measure between two nodes. These methods generally propose a strategy based on random walks on graphs (i. e., the interaction network). A random walk of a certain length k on a graph represents a stochastic process starting at a seed node and each subsequently visited node is chosen uniformly at random from the neighbors of each previous node. The steady state probability $p_{x,y,k}(G)$ is then the probability that a random walk of length k , starting at node x would end in node y .

Kohler et al. [19.95] proposed a variant of the random walk, namely the random walk with restart, to identify disease genes in a human PPI network. Here, the walk is allowed to restart in every time step at a known disease gene (i. e., seed gene) with a certain probability. All candidate disease genes are then ranked according to the probability of the random walker reaching the candidate gene from a (set of) seed node(s), which reflects their global relatedness to the known disease seed genes. Applied on 110 disease gene families, together with their known associated genes in a leave-one-out cross-validation setting, they clearly outperformed local measures as direct neighborhood or shortest path, or approaches not based on any network.

Vanunu et al. [19.59] extended this approach by using causal genes both from the disease of interest or similar ones as seed genes in the random walk with restart. This approach can be very useful when no causal genes of the disease are known. Diseases, however, can be very heterogeneous, meaning that they can result in different phenotypes and encompass various subtypes. Exploiting all known disease genes that have been related to a heterogeneous disease might not have a sufficiently high resolution to predict novel genes for a specific subtype of the disease. To overcome this, two other methods proposed to identify gene-phenotype relationships rather than finding the gene-disease relationship directly. This strategy decomposes a disease in phenotypes and tries to identify novel phenotype-related genes by using genes related to one or more phenotypes of interest and related ones as seed genes. *Li and Patra* [19.96] build upon the random walk with restart strategy, but perform it on a heterogeneous network connecting the gene network (i. e., consisting of PPI) with the phenotype network (i. e., a k nearest neighbor graph presentation of phenotypes similarity) through gene-phenotype relations. *Yang et al.* [19.97], on the other hand, make use of the associations between protein complexes and the phenotypes of interest to perform a random walk with restart on a protein complex network.

Case Study

The aforementioned methods rank candidate genes based on their connections to known disease genes or to known causal genes of related phenotypes in a protein–protein network. However, these methods are usually ineffective when little is known about the molecular basis of the phenotype (e. g., no confirmed causal genes), or when the observed phenotypes are very specific. To this end, *Nitsch et al.* [19.98] developed a computational method to prioritize candidate disease genes for which limited or no prior knowledge is available, by using experimental data on differential gene expression between healthy and affected individuals for a phenotype of interest.

Genes for which significant differential expression was measured in an affected tissue compared to wild type are usually considered as promising candidates being involved in the disease. However, not necessarily the expression levels of the disease gene are affected, but rather expression of genes downstream of this causal gene. Therefore, by mapping differential expression levels on a gene network, one expects to observe a disrupted expression module around the dis-

ease gene. Other candidate genes that are not causally related to the phenotype should not be part of such a module. For this reason, the relevance of a candidate gene is scored by considering the level of differential expression in its neighborhood in a protein network instead of only taking its own expression level into account, under the assumption that strong candidates tend to be surrounded by differentially expressed neighbors.

In this work, a functional network was created using human protein associations obtained from the **STRING** database [19.26], since protein interaction networks are still far from complete and according to the authors might give suboptimal results due to many missing components and pathways. For each gene, differential expression values are determined from microarray experiments that measure wild type versus diseased cell lines, for a phenotype of interest. The prioritization can be performed on a list of candidate genes on a chromosomal region of interest (e.g., determined from a linkage study) or genome-wide when no list of candidate genes are available (although this will probably result in many more false positives).

The neighborhood of a candidate gene is then determined by using a graph kernel, namely the Laplacian exponential diffusion kernel. This gives a weight to each gene in the network, which decreases as a function of the distance from the causal gene, taking into account that there might be multiple paths between the causal gene and each gene in the network. It can be seen as a random walk, starting from a node and transitioning to a neighboring node with a certain probability.

19.4 Conclusion

Despite the emerging amount of data and the integration strategies presented at the beginning of this chapter, eukaryotic molecular networks are still low in coverage due to their size and complexity. We expect that the increasing number of available data sets will continue to expand current networks and further refine our knowledge on the included interactions, ultimately resulting in a better understanding of the cell's behavior. In this

chapter we have presented several path finding methodologies to interrogate networks with functional data and showed how they can be used to predict disease genes, unveil hidden signaling paths, . . . The use of interaction networks for unveiling mode of actions will only gain in importance with functional omics datasets profiling human diseases growing steadily in number and size.

Finally, each candidate gene is scored by summing up the levels of differential expression (measures by absolute fold changes) of each gene, weighted by its network distance from the candidate. Higher differential expression of neighboring genes will, therefore, result in higher scores. The significance of a candidate gene is determined by randomly distributing the differential expression data on the network and computing an empirical p-value from the random distribution of scores. Besides benchmarking, the methodology on several monogenic diseases for which the causal gene is known, the authors also applied their method on the polygenic disorder *Stein–Leventhal* [19.100] for which currently no disease gene is known. They highly ranked two genes on two different chromosomal regions that were previously assigned a possible role in this disorder, namely fibrillin 3 (**FBN3**) and follistatin (**FST**). Another gene, **DEAD box4**, was found to be the best scoring gene and was suggested as a new candidate gene potentially involved in this disease. Although little is known about the molecular function of **DEAD box 4** in mammals, the authors found several indications in literature that indicate a plausible role in the *Stein–Leventhal* syndrome, for instance because of its association with stem cell recruitment to the ovaries, interaction with the **mRNA** processing machinery, and impact on apoptosis.

Although the expression levels of genes can be determined by multiple genes together in the case of a polygenic order, making it difficult to determine the true causes of the effects, the approach has nevertheless shown to provide plausible candidates when little knowledge is available for the disease at hand.

References

- 19.1 U. Alon: Biological networks: The tinkerer as an engineer, *Science* **301**(5641), 1866–1867 (2003)
- 19.2 R. Milo, S. Shen–Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon: Network motifs: Simple building blocks of complex networks, *Science* **298**(5594), 824–827 (2002)
- 19.3 U. Alon: Network motifs: Theory and experimental approaches, *Nat. Rev. Genet.* **8**(6), 450–461 (2007)

- 19.4 Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutillier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreau, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A.R. Willems, H. Sassi, P.A. Nielsen, K.J. Rasmussen, J.R. Andersen, L.E. Johansen, L.H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B.D. Sørensen, J. Matthiesen, R.C. Hendrickson, F. Gleeson, T. Pawson, M.F. Moran, D. Durocher, M. Mann, C.W. Hogue, D. Figeys, M. Tyers: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature* **415**(6868), 180–183 (2002)
- 19.5 T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki: A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc. Natl. Acad. Sci. USA* **98**(8), 4569–4574 (2001)
- 19.6 N.J. Krogan, G. Cagney, H. Yu, G. Zhong, X. Guo, A. Ignatchenko, J. Li, S. Pu, N. Datta, A.P. Tikuisis, T. Punna, J.M. Peregrín-Alvarez, M. Shales, X. Zhang, M. Davey, M.D. Robinson, A. Paccanaro, J.E. Bray, A. Sheung, B. Beattie, D.P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M.M. Canete, J. Vlasblom, S. Wu, C. Orsi, S.R. Collins, S. Chandran, R. Haw, J.J. Rilstone, K. Gandi, N.J. Thompson, G. Musso, P. St. Onge, S. Ghanny, M.H. Lam, G. Butland, A.M. Altaf-Ul, S. Kanaya, A. Shilatifard, E. O’Shea, J.S. Weissman, C.J. Ingles, T.R. Hughes, J. Parkinson, M. Gerstein, S.J. Wodak, A. Emili, J.F. Greenblatt: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*, *Nature* **440**(7084), 637–643 (2006)
- 19.7 P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamar, M. Yang, M. Johnston, S. Fields, J.M. Rothberg: A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*, *Nature* **403**(6770), 623–627 (2000)
- 19.8 A.C. Gavin, M. Bösch, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J.M. Rick, A.M. Michon, C.M. Cruciat, M. Remor, C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.A. Heurtier, R.R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, G. Superti-Furga: Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature* **415**(6868), 141–147 (2002)
- 19.9 S. Li, C.M. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.O. Vidalain, J.D. Han, A. Chesneau, T. Hao, D.S. Goldberg, N. Li, M. Martinez, J.F. Rual, P. Lamesch, L. Xu, M. Tewari, S.L. Wong, L.V. Zhang, G.F. Berriz, L. Jacotot, P. Vaglio, J. Reboul, T. Hirozane-Kishikawa, Q. Li, H.W. Gabel, A. Elewa, B. Baumgartner, D.J. Rose, H. Yu, S. Bosak, R. Sequerra, A. Fraser, S.E. Mango, W.M. Saxton, S. Strome, S. van den Heuvel, F. Piano, J. Vandenhaute, C. Sardet, M. Gerstein, L. Doucette-Stamm, K.C. Gunsalus, J.W. Harper, M.E. Cusick, F.P. Roth, D.E. Hill, M. Vidal: A map of the interactome network of the metazoan *C. elegans*, *Science* **303**(5657), 540–543 (2004)
- 19.10 E. Formstecher, S. Aresta, V. Collura, A. Hamburger, A. Meil, A. Trehin, C. Reverdy, V. Betin, S. Maire, C. Brun, B. Jacq, M. Arpin, Y. Bellaiche, S. Bellusci, P. Benaroch, M. Bornens, R. Chanet, P. Chavrier, O. Delattre, V. Doye, R. Fehon, G. Faye, T. Galli, J.A. Girault, B. Goud, J. de Gunzburg, L. Johannes, M.P. Junier, V. Mirouse, A. Mukherjee, D. Papadopoulou, F. Perez, A. Plessis, C. Rossé, S. Saule, D. Stoppa-Lyonnet, A. Vincent, M. White, P. Legrain, J. Wojcik, J. Camonis, L. Daviet: Protein interaction mapping: A *Drosophila* case study, *Genome Res.* **15**(3), 376–384 (2005)
- 19.11 L. Giot, J.S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y.L. Hao, C.E. Ooi, B. Godwin, E. Vitols, G. Vijayadamar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C.A. Stanyon, R.L. Finley Jr., K.P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R.A. Shimkets, M.P. McKenna, J. Chant, J.M. Rothberg: A protein interaction map of *Drosophila melanogaster*, *Science* **302**(5651), 1727–1736 (2003)
- 19.12 J.F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G.F. Berriz, F.D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D.S. Goldberg, L.V. Zhang, S.L. Wong, G. Franklin, S. Li, J.S. Albala, J. Lim, C. Fraughton, E. Llamasas, S. Cevik, C. Bex, P. Lamesch, R.S. Sikorski, J. Vandenhaute, H.Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M.E. Cusick, D.E. Hill, F.P. Roth, M. Vidal: Towards a proteome-scale map of the human protein–protein interaction network, *Nature* **437**(7062), 1173–1178 (2005)
- 19.13 U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F.H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, E.E. Wanker: A human protein–protein interaction network:

- A resource for annotating the proteome, *Cell* **122**(6), 957–968 (2005)
- 19.14 T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, R.A. Young: Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science* **298**(5594), 799–804 (2002)
- 19.15 C.T. Harbison, D.B. Gordon, T.I. Lee, N.J. Rinaldi, K.D. Macisaac, T.W. Danford, N.M. Hannett, J.B. Tagne, D.B. Reynolds, J. Yoo, E.G. Jennings, J. Zeitlinger, D.K. Pokholok, M. Kellis, P.A. Rolfe, K.T. Takusagawa, E.S. Lander, D.K. Gifford, E. Fraenkel, R.A. Young: Transcriptional regulatory code of a eukaryotic genome, *Nature* **431**(7004), 99–104 (2004)
- 19.16 S.E. Celniker, L.A. Dillon, M.B. Gerstein, K.C. Gunsalus, S. Henikoff, G.H. Karpen, M. Kellis, E.C. Lai, J.D. Lieb, D.M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K.P. White, R.H. Waterston, modENCODE Consortium: Unlocking the secrets of the genome, *Nature* **459**(7249), 927–930 (2009)
- 19.17 R. Aebersold, M. Mann: Mass spectrometry-based proteomics, *Nature* **422**(6928), 198–207 (2003)
- 19.18 M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa: KEGG for representation and analysis of molecular networks involving diseases and drugs, *Nucleic Acids Res.* **38**(Database issue), D355–360 (2010)
- 19.19 R. Caspi, T. Altman, J.M. Dale, K. Dreher, C.A. Fulcher, F. Gilham, P. Kaipa, A.S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, S. Paley, L. Popescu, A. Pujar, A.G. Shearer, P. Zhang, P.D. Karp: The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic Acids Res.* **38**(Database issue), D473–479 (2010)
- 19.20 J. Schellenberger, J.O. Park, T.M. Conrad, B.O. Palsson: BiGG: A biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions, *BMC Bioinformatics* **11**, 213 (2010)
- 19.21 C. Stark, B.J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M.S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, T. Reguly, J.M. Rust, A. Winter, K. Dolinski, M. Tyers: The BioGRID interaction database: 2011 update, *Nucleic Acids Res.* **39**(Database issue), D698–704 (2011)
- 19.22 G.D. Bader, D. Betel, C.W. Hogue: BIND: The biomolecular interaction network database, *Nucleic Acids Res.* **31**(1), 248–250 (2003)
- 19.23 I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, D. Eisenberg: DIP, the Database of interacting proteins: A research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.* **30**(1), 303–305 (2002)
- 19.24 A. Ceol, A. Chatr-Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, G. Cesareni: MINT, the molecular interaction database: 2009 update, *Nucleic Acids Res.* **38**(Database issue), D532–539 (2010)
- 19.25 H.W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K.F. Mayer, V. Stumpflen, A. Antonov: MIPS: Curated databases and comprehensive secondary data resources in 2010, *Nucleic Acids Res.* **39**(Database issue), D220–224 (2011)
- 19.26 C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, B. Snel: STRING: A database of predicted functional associations between proteins, *Nucleic Acids Res.* **31**(1), 258–261 (2003)
- 19.27 R. Goel, B. Muthusamy, A. Pandey, T.S. Prasad: Human protein reference database and human proteinpedia as discovery resources for molecular biotechnology, *Mol. Biotechnol.* **48**(1), 87–95 (2011)
- 19.28 A. Ruepp, B. Waegle, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, H.W. Mewes: CORUM: The comprehensive resource of mammalian protein complexes – 2009, *Nucleic Acids Res.* **38**(Database issue), D497–501 (2010)
- 19.29 E. Wingender: The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation, *Brief Bioinf.* **9**(4), 326–332 (2008)
- 19.30 P.E. Hodges, A.H. McKee, B.P. Davis, W.E. Payne, J.I. Garrels: The yeast proteome database (YPD): A model for the organization and presentation of genome-wide functional data, *Nucleic Acids Res.* **27**(1), 69–73 (1999)
- 19.31 C.Y. Yang, C.H. Chang, Y.L. Yu, T.C. Lin, S.A. Lee, C.C. Yen, J.M. Yang, J.M. Lai, Y.R. Hong, T.L. Tseng, K.M. Chao, C.Y. Huang: PhosphoPOINT: A comprehensive human kinase interactome and phospho-protein database, *Bioinformatics* **24**(16), i14–20 (2008)
- 19.32 P.V. Hornbeck, I. Chabra, J.M. Kornhauser, E. Skrzypek, B. Zhang: PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation, *Proteomics* **4**(6), 1551–1561 (2004)
- 19.33 G.R. Cochrane, M.Y. Galperin: The 2010 nucleic acids research database issue and online database collection: A community of data resources, *Nucleic Acids Res.* **38**(Database issue), D1–4 (2010)
- 19.34 G.D. Bader, M.P. Cary, C. Sander: Pathguide: A pathway resource list, *Nucleic Acids Res.* **34**(Database issue), D504–506 (2006)
- 19.35 A. Beyer, C. Workman, J. Hollunder, D. Radke, U. Moller, T. Wilhelm, T. Ideker: Integrated assessment and prediction of transcription factor binding, *PLoS Comput. Biol.* **2**(6), e70 (2006)
- 19.36 G.T. Hart, A.K. Ramani, E.M. Marcotte: How complete are current yeast and human protein–protein

- interaction networks?, *Genome Biol.* **7**(11), 120 (2006)
- 19.37 F. Ramirez, A. Schlicker, Y. Assenov, T. Lengauer, M. Albrecht: Computational analysis of human protein interaction networks, *Proteomics* **7**(15), 2541–2552 (2007)
- 19.38 C. von Mering, R. Krause, B. Snel, M. Cornell, S.G. Oliver, S. Fields, P. Bork: Comparative assessment of large-scale data sets of protein–protein interactions, *Nature* **417**(6887), 399–403 (2002)
- 19.39 C.M. Bishop: *Pattern Recognition and Machine Learning* (Springer, New York 2006)
- 19.40 D. Ucar, A. Beyer, S. Parthasarathy, C.T. Workman: Predicting functionality of protein–DNA interactions by integrating diverse evidence, *Bioinformatics* **25**(12), i137–i144 (2009)
- 19.41 J.S. Bader, A. Chaudhuri, J.M. Rothberg, J. Chant: Gaining confidence in high-throughput protein interaction networks, *Nat. Biotechnol.* **22**(1), 78–85 (2004)
- 19.42 R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, M. Gerstein: A Bayesian networks approach for predicting protein–protein interactions from genomic data, *Science* **302**(5644), 449–453 (2003)
- 19.43 D.R. Rhodes, S.A. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana–Sundaram, D. Ghosh, A. Pandey, A.M. Chinnaiyan: Probabilistic model of the human protein–protein interaction network, *Nat. Biotechnol.* **23**(8), 951–959 (2005)
- 19.44 M.S. Scott, G.J. Barton: Probabilistic prediction and ranking of human protein–protein interactions, *BMC Bioinformatics* **8**, 239 (2007)
- 19.45 R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, T. Ideker: Conserved patterns of protein interaction in multiple species, *Proc. Natl. Acad. Sci. USA* **102**(6), 1974–1979 (2005)
- 19.46 M.A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G.M. Rubin, J.A. Blake, C. Bult, M. Dolan, H. Drabkin, J.T. Eppig, D.P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J.M. Cherry, K.R. Christie, M.C. Costanzo, S.S. Dwight, S. Engel, D.G. Fisk, J.E. Hirschman, E.L. Hong, R.S. Nash, A. Sethuraman, C.L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S.Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E.M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, Gene Ontology Consortium: The gene ontology (GO) database and informatics resource, *Nucleic Acids Res.* **32**(Database issue), D258–261 (2004)
- 19.47 Y. Guan, C.L. Myers, R. Lu, I.R. Lemischka, C.J. Bult, O.G. Troyanskaya: A genomewide functional network for the laboratory mouse, *PLoS Comput. Biol.* **4**(9), e1000165 (2008)
- 19.48 T. Shlomi, D. Segal, E. Ruppin, R. Sharan: QPath: A method for querying pathways in a protein–protein interaction network, *BMC Bioinformatics* **7**, 199 (2006)
- 19.49 I. Lee, S.V. Date, A.T. Adai, E.M. Marcotte: A probabilistic functional network of yeast genes, *Science* **306**(5701), 1555–1558 (2004)
- 19.50 I. Lee, Z. Li, E.M. Marcotte: An improved, bias-reduced probabilistic functional gene network of baker’s yeast, *Saccharomyces cerevisiae*, *PLoS One* **2**(10), e988 (2007)
- 19.51 I. Lee, B. Lehner, C. Crombie, W. Wong, A.G. Fraser, E.M. Marcotte: A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*, *Nat. Genet.* **40**(2), 181–188 (2008)
- 19.52 I. Lee, B. Ambaru, P. Thakkar, E.M. Marcotte, S.Y. Rhee: Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*, *Nat. Biotechnol.* **28**(2), 149–156 (2010)
- 19.53 B. Linghu, E.S. Snitkin, Z. Hu, Y. Xia, C. Delisi: Genome-wide prioritization of disease genes and identification of disease–disease associations from an integrated human functional linkage network, *Genome Biol.* **10**(9), R91 (2009)
- 19.54 I. Lee, U.M. Blom, P.I. Wang, J.E. Shim, E.M. Marcotte: Prioritizing candidate disease genes by network-based boosting of genome-wide association data, *Genome Res.* **21**(7), 1109–1121 (2011)
- 19.55 J. Scott, T. Ideker, R.M. Karp, R. Sharan: Efficient algorithms for detecting signaling pathways in protein interaction networks, *J. Comput. Biol.* **13**(2), 133–144 (2006)
- 19.56 G. Bebek, J. Yang: PathFinder: Mining signal transduction pathway segments from protein–protein interaction networks, *BMC Bioinformatics* **8**, 335 (2007)
- 19.57 O. Ourfali, T. Shlomi, T. Ideker, E. Ruppin, R. Sharan: SPINE: A framework for signaling–regulatory pathway inference from cause–effect experiments, *Bioinformatics* **23**(13), i359–366 (2007)
- 19.58 N. Yosef, L. Ungar, E. Zalckvar, A. Kimchi, M. Kupiec, E. Ruppin, R. Sharan: Toward accurate reconstruction of functional protein networks, *Mol. Syst. Biol.* **5**, 248 (2009)
- 19.59 O. Vanunu, O. Mager, E. Ruppin, T. Shlomi, R. Sharan: Associating genes and protein complexes with disease via network propagation, *PLoS Comput. Biol.* **6**(1), e1000641 (2010)
- 19.60 Y. Qi, Z. Bar–Joseph, J. Klein–Seetharaman: Evaluation of different biological data and computational classification methods for use in protein interaction prediction, *Proteins* **63**(3), 490–500 (2006)
- 19.61 S. Suthram, T. Shlomi, E. Ruppin, R. Sharan, T. Ideker: A direct comparison of protein in-

- teraction confidence assignment schemes, *BMC Bioinformatics* **7**, 360 (2006)
- 19.62 Y. Qi, J. Klein-Seetharaman, Z. Bar-Joseph: Random forest similarity for protein-protein interaction prediction from multiple sources, *Pac. Symp. Biocomput.* (2005) pp. 531–542
- 19.63 X. Zhu, M. Gerstein, M. Snyder: Getting connected: Analysis and principles of biological networks, *Genes Dev.* **21**(9), 1010–1024 (2007)
- 19.64 E. Yeger-Lotem, L. Riva, L.J. Su, A.D. Gitler, A.G. Cashikar, O.D. King, P.K. Auluck, M.L. Geddie, J.S. Valastyan, D.R. Karger, S. Lindquist, E. Fraenkel: Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity, *Nat. Genet.* **41**(3), 316–323 (2009)
- 19.65 E. Cerami, E. Demir, N. Schultz, B.S. Taylor, C. Sander: Automated network analysis identifies core pathways in glioblastoma, *PLoS One* **5**(2), e8918 (2010)
- 19.66 J. Bellay, G. Atluri, T.L. Sing, K. Toufighi, M. Costanzo, P.S. Ribeiro, G. Pandey, J. Baller, B. VanderSluis, M. Michaut, S. Han, P. Kim, G.W. Brown, B.J. Andrews, C. Boone, V. Kumar, C.L. Myers: Putting genetic interactions in context through a global modular decomposition, *Genome Res.* **21**(8), 1375–1387 (2011)
- 19.67 C.Y. Huang, E. Fraenkel: Integration of proteomic, transcriptional, and interactome data reveals hidden signaling components, *Sci. Signal.* **2**, ra40 (2009)
- 19.68 X.M. Zhao, R.S. Wang, L. Chen, K. Aihara: Uncovering signal transduction networks from high-throughput data by integer linear programming, *Nucleic Acids Res.* **36**(9), e48 (2008)
- 19.69 S. Lu, F. Zhang, J. Chen, S. Sze: Finding pathway structures in protein interaction networks, *Algoritmica* **48**, 363–374 (2007)
- 19.70 X. Ren, X. Zhou, L.Y. Wu, X.S. Zhang: An information-flow-based model with dissipation, saturation and direction for active pathway inference, *BMC Syst. Biol.* **4**, 72 (2010)
- 19.71 M. Steffen, A. Petti, J. Aach, P. D’Haeseleer, G. Church: Automated modelling of signal transduction networks, *BMC Bioinformatics* **3**, 34 (2002)
- 19.72 A. Joshi, T. Van Parys, Y.V. Peer, T. Michoel: Characterizing regulatory path motifs in integrated networks using perturbational data, *Genome Biol.* **11**(3), R32 (2010)
- 19.73 J. Pey, J. Prada, J.E. Beasley, F.J. Planes: Path finding methods accounting for stoichiometry in metabolic networks, *Genome Biol.* **12**(5), R49 (2011)
- 19.74 N. Novershtern, A. Regev, N. Friedman: Physical Module Networks: An integrative approach for reconstructing transcription regulation, *Bioinformatics* **27**(13), i177–i185 (2011)
- 19.75 Z. Hu, P.J. Killion, V.R. Iyer: Genetic reconstruction of a functional transcriptional regulatory network, *Nat. Genet.* **39**(5), 683–687 (2007)
- 19.76 G. Chua, Q.D. Morris, R. Sopko, M.D. Robinson, O. Ryan, E.T. Chan, B.J. Frey, B.J. Andrews, C. Boone, T.R. Hughes: Identifying transcription factor functions and targets by phenotypic activation, *Proc. Natl. Acad. Sci. USA* **103**(32), 12045–12050 (2006)
- 19.77 S.S. Shen-Orr, R. Milo, S. Mangan, U. Alon: Network motifs in the transcriptional regulation network of *Escherichia coli*, *Nat. Genet.* **31**(1), 64–68 (2002)
- 19.78 L.V. Zhang, O.D. King, S.L. Wong, D.S. Goldberg, A.H. Tong, G. Lesage, B. Andrews, H. Bussey, C. Boone, F.P. Roth: Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network, *J. Biol.* **4**(2), 6 (2005)
- 19.79 N.M. Luscombe, M.M. Babu, H. Yu, M. Snyder, S.A. Teichmann, M. Gerstein: Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature* **431**(7006), 308–312 (2004)
- 19.80 C.H. Yeang, T. Ideker, T. Jaakkola: Physical network models, *J. Comput. Biol.* **11**(2/3), 243–262 (2004)
- 19.81 C.H. Yeang, M. Vingron: A joint model of regulatory and metabolic networks, *BMC Bioinformatics* **7**, 332 (2006)
- 19.82 A. Medvedovsky, V. Bafna, U. Zwick, R. Sharan: An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks, *LNC5* **5251**, 222–232 (2008)
- 19.83 A. Gitter, J. Klein-Seetharaman, A. Gupta, Z. Bar-Joseph: Discovering pathways by orienting edges in protein interaction networks, *Nucleic Acids Res.* **39**(4), e22 (2011)
- 19.84 T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburty, J. Simon, M. Bard, S.H. Friend: Functional discovery via a compendium of expression profiles, *Cell* **102**(1), 109–126 (2000)
- 19.85 Y. Gilad, S.A. Rifkin, J.K. Pritchard: Revealing the architecture of gene regulation: The promise of eQTL studies, *Trends Genet.* **24**(8), 408–415 (2008)
- 19.86 S. Suthram, A. Beyer, R.M. Karp, Y. Eldar, T. Ideker: eQED: An efficient method for interpreting eQTL associations using protein networks, *Mol. Syst. Biol.* **4**, 162 (2008)
- 19.87 Z. Tu, L. Wang, M.N. Arbeitman, T. Chen, F. Sun: An integrative approach for causal gene identification and gene regulatory pathway inference, *Bioinformatics* **22**(14), e489–496 (2006)
- 19.88 R.B. Brem, G. Yvert, R. Clinton, L. Kruglyak: Genetic dissection of transcriptional regulation in budding yeast, *Science* **296**(5568), 752–755 (2002)
- 19.89 G. Yvert, R.B. Brem, J. Whittle, J.M. Akey, E. Foss, E.N. Smith, R. Mackelprang, L. Kruglyak:

- Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors, *Nat. Genet.* **35**(1), 57–64 (2003)
- 19.90 Y.A. Kim, S. Wuchty, T.M. Przytycka: Identifying causal genes and dysregulated pathways in complex diseases, *PLoS Comput. Biol.* **7**(3), e1001095 (2011)
- 19.91 J. Amberger, C.A. Bocchini, A.F. Scott, A. Hamosh: McKusick's online Mendelian inheritance in man (OMIM), *Nucleic Acids Res.* **37**(Database issue), D793–796 (2009)
- 19.92 M. Oti, H.G. Brunner: The modular nature of genetic diseases, *Clin. Genet.* **71**(1), 1–11 (2007)
- 19.93 M. Oti, B. Snel, M.A. Huynen, H.G. Brunner: Predicting disease genes using protein-protein interactions, *J. Med. Genet.* **43**(8), 691–698 (2006)
- 19.94 R.A. George, J.Y. Liu, L.L. Feng, R.J. Bryson-Richardson, D. Fatkin, M.A. Wouters: Analysis of protein sequence and interaction data for candidate disease gene prediction, *Nucleic Acids Res.* **34**(19), e130 (2006)
- 19.95 S. Kohler, S. Bauer, D. Horn, P.N. Robinson: Walking the interactome for prioritization of candidate disease genes, *Am. J. Hum. Genet.* **82**(4), 949–958 (2008)
- 19.96 Y. Li, J.C. Patra: Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network, *Bioinformatics* **26**(9), 1219–1224 (2010)
- 19.97 P. Yang, X. Li, M. Wu, C.K. Kwoh, S.K. Ng: Inferring gene-phenotype associations via global protein complex network propagation, *PLoS One* **6**(7), e21502 (2011)
- 19.98 D. Nitsch, L.C. Tranchevent, B. Thienpont, L. Thorrez, H. Van Esch, K. Devriendt, Y. Moreau: Network analysis of differential expression for the identification of disease-causing genes, *PLoS One* **4**(5), e5526 (2009)
- 19.99 C. Boone, H. Bussey, B.J. Andrews: Exploring genetic interactions and networks with yeast, *Nat. Rev. Genet.* **8**(6), 437–449 (2007)
- 19.100 M. Cortón, J.I. Botella-Carretero, A. Benguria, G. Villuendas, A. Zaballos, J.L. San Millan, H.F. Escobar-Morreale, B. Peral: Differential gene expression profile in omental adipose tissue in women with polycystic ovary syndrome, *J. Clin. Endocrinol. Metab.* **92**(1), 328–337 (2007)