Molecular Networks - Representation and Analysis

Miguel A. Hernández-Prieto, Ravi Kiran Reddy Kalathur & Matthias E. Futschik

Centro de Biomedicina Molecular e Estrutural, Universidade do Algarve, Faro, Portugal

The original publication is available at www.springerlink.com.

Summary

Molecular networks, their representation and analysis have attracted increasing interest in recent years. Although the importance of molecular networks has been recognized for a long time, only the advent of new technologies during the last two decades have delivered the necessary data for a systematic study of molecular networks and their complex behavior. Especially the surge of genome-wide data as well as the increase in computational power have contributed establishing network and systems biology as new paradigms. The conceptual framework is generally based on an integrated approach of computational and experimental methods. In this chapter, we introduce basic concepts and outline mathematical formalisms for representing and analyzing molecular networks. In particular, we review the study of transcriptional regulatory networks in prokaryotes and of protein interaction networks in human as prime examples for network-orientated approaches to complex systems. The chapter is concluded with a discussion of current challenges and future directions of network biology.

1 Overview

Biological systems can range from a tiny bacterium in the soil to large food webs across the oceans. Despite striking differences in size and appearance, they have one thing in common: they are complex networks made out of numerous single components interacting with each other. The cell can be seen as a leading example of a highly organized network with a plethora of single parts miraculously interwoven. While biological networks are truly fascinating, they are at the same time notoriously difficult to study due to their intrinsic complexity. For a long time, this was especially the case for molecular networks, as lack of technology limited the simultaneous measurement of the various components and interactions. This situation has dramatically changed with technological breakthroughs in genomics, proteomics and metabolomics enabling the profiling of thousands of molecules.

In molecular biology, this had motivated a shift of paradigms from reductionism, that places single components in the foreground, to systems biology, where we aim to study molecular processes as a whole [1-2]. The key idea of systems biology is that the total system is more than the simple sum of its single components, i.e. new properties of a system are emerging that are not exhibited by its single

isolated components. The two main branches of systems biology are (i) generation of systems-wide *in vitro* and *in vivo* (and possibly quantitative) data and (ii) *in silico* representation and analysis of the biological systems. Besides providing a better understanding of complex molecular mechanisms, systems biology aims to establish a framework for integration and consolidation of different types of data. This is particular important, as the current bottleneck in biomedical research is frequently not the generation of data, but their analysis and interpretation.

The application of systems biology can be manifold. Signaling pathways provide an excellent case how models of complex mechanisms can help biomedical research. The identification of one disease target in a pathway implies that other components of the same pathway may serve as alternative drug targets increasing drastically the possibilities for novel therapeutic interventions. For instance, the cholesterol synthesis pathway, whose end product is linked to cardiovascular diseases, can be nowadays targeted by multiple drugs. It should be however noted, that although the term systems biology has been coined only recently, many ideas of systems biology have been introduced much earlier.

A major challenge of systems biology is the adequate representation of the system. Numerous components might have to be included for a consistent model. Especially for larger biological systems, the representation and analysis is computationally challenging. Here, concepts of network biology might help [3]. Its basis is the representation of biological systems as mathematical graphs. This leads to considerable simplification (see figure 1 as example): Molecular networks are conceived as connected sets of vertices, which can be analyzed using well established tools of graph theory.

The importance of network-based approaches and their implementation will be illustrated on two specific types of networks in this chapter: First, we will review the study of transcriptional regulatory networks in prokaryotes. Transcriptional regulation in prokaryotes has been a focus of molecular biology, since the seminal work by Monod and Jacobs elucidating the genetic regulation of the Lac operon [4]. In fact, many insights into molecular networks and their functioning have been gained by the study of transcriptional regulation in prokaryotes. We will thus introduce the underlying biology and give an overview how to model regulatory networks.

As the second type of networks, physical protein interaction networks in human will be discussed. Proteins are of crucial importance for the correct functioning and orchestration of many cellular processes. Most proteins do not function alone, but within a cellular context network by interactions with other proteins. To obtain a better understanding, several large-scale interaction networks for human proteins have been constructed. Besides summarizing current approaches to obtain a human protein interactome, we present some network-based studies in cancer research. Especially for cancer, systems and network biology might provide promising tools to deal with the variability and heterogeneity of carcinogenesis.

It should be noted that other type of molecular networks might allow - or even demand - different kind of representations and analysis tools. For instance, detailed knowledge of metabolic pathways has enabled the development of advanced mathematical frameworks, which can be used for a quantitative prediction instead of solely qualitative description [5]. For the types of networks, which we will be discussed here, the availability of high quality data is more restricted. Also, regulatory transcriptional networks and proteins interaction networks exhibit a high degree of intrinsic complexity making the formulation of exact models challenging.

2 Transcriptional Regulatory Networks in Prokaryotes

Transcriptional regulatory networks have drawn considerable attention because of various reasons. Besides the experimental ease of studying prokaryotes in the laboratory and their general importance in biology and medicine, prokaryotes are prime examples for adaption and optimization of intrinsic cellular mechanisms to extrinsic conditions. Due to evolutionary pressure, the genomes of the first prokaryotic organisms have been extensively re-modeled to adapt to different environmental conditions. This adaptation is facilitated by horizontal gene transfer and a fast generation time [6]. As consequence, prokaryotic organisms have successfully conquered most ecological niches on earth. Notably, the increasing number of sequenced genomes as well as of genome-wide expression studies has facilitated the study of prokaryotic adaption by means of bioinformatics and systems biology.

Before we discuss in more detail approaches to elucidate transcriptional regulatory networks, we briefly review the main components of prokaryotic transcriptional regulation in the next sections.

2.1 Transcription in Prokaryotes

Transcription describes the process by which RNA polymerases read and copy information from DNA to RNA. The prokaryotic RNA polymerase is a complex formed by several catalytic subunits and a single regulatory subunit known as sigma (σ) factor. Transcription can be generally divided in four phases: (i) pre-initiation, (ii) initiation, (iii) elongation and (iv) termination. For regulation of gene expression, pre-initiation and initiation are the most important phases.

During pre-initiation, a sigma factor binds to an upstream sequence region (also called the promoter region) of the gene to be transcribed. Usually, several sigma factors exist in bacterial genomes. One is normally considered as principal sigma factor responsible for expression of housekeeping genes, while the other sigma factors will regulate expression of genes required under specific conditions. After sigma factor binding, the RNA polymerase associates with promoter elements situated approximately at -10 and -35 bases upstream of the transcription start site.

Besides sequences recognized by sigma factors, the promoter can contain other specific sites which determine binding of different regulatory elements. In particular, short sequences can act as binding motifs for transcription factors which can be activators or repressors of gene expression. Thus, both sigma and transcription factors regulators recognize specific DNA patterns in promoter regions; and it is the combinations of these sequence elements that results in selectivity of gene expression. It is important to note that the initiation of transcription is not ensured by binding of sigma or transcription factors to the promoter, as cells have a limited pool of RNA polymerases and therefore subsequent assembly of RNA polymerases is rather a competitive process.

After binding of the RNA polymerase to the sigma factor and initiation of transcription, elongation continues until the RNA polymerase core releases the RNA transcript and dissociates from the DNA template. The dissociation can be provoked by binding of a so-called rho co-factor or formation of hairpin structures in rho-independent transcriptional terminators [7]. The synthesized RNA can then serve as template (in form of messenger RNA) for protein synthesis or can have functionality by itself (e.g. in form of ribosomal RNA).

2.2 Transcription Factors and Their Binding Sites

As mentioned earlier, transcription factors are the main determinants for transcription initiation besides sigma factors. Their binding to specific sequences in the promoter region is typically regulated in response to intracellular or environmental stimuli. The transcription factor binding sites are normally 12 to 30 nucleotides long. To understand specificity of transcriptions, it is important to identify such binding sites. A common way is to search for a conserved short sequence (i.e. a consensus sequence) in the promoter region of genes under the control of the transcription factor. Alternatively, one can determine the binding site experimentally. To this end, the promoter region is fused to a reporter gene and subsequently truncated or modified until the area responsible for the specific gene regulation is localized. Mutation analysis can then be used to confirm the transcription factor binding site.

Frequently, our primary interest does not lay on regulation of a single gene but on the characterization of genome-wide effects of transcription factors. In this case, chromatin immunoprecipitation can be applied in case a specific antibody for the transcription factor is available. The technique permits the co-isolation of the transcription factor and its bound DNA. Subsequently, the isolated double stranded DNA can then be labeled and hybridized onto a microarray (ChIP-chip) [8-10] or sequenced (ChIP-seq) [11-12] in order to determine the promoters which were bound to the transcription factor.

2.3 Gene Expression Control by Small RNAs

Only recently, the importance of small non-coding RNAs (ncRNAs) for regulation of gene expression has been fully appreciated, despite evidence for regulatory ncRNA in bacteria was reported already over 30 years ago [13-14]. The list of ncRNAs detected in prokaryotes has rapidly increased with the number of new genomes being sequenced. Approaches for their identification range from computational predictions [15] to experimental techniques such as high-throughput pyro-sequencing [16]. ncRNAs can influence gene expression in various ways. They can bind directly to specific mRNAs affecting their stability or translation rate; or they can attach to proteins modifying their activity. Although the function of many ncRNAs is still unknown, their importance in gene regulation has become undisputed. In fact, cellular response to environmental changes seems to frequently depend of a tight coordination between ncRNAs and protein regulators. One intriguing example is the interplay of transcription factor Fur (Ferric uptake regulator) and an ncRNA termed RhyB in order to ensure fast response to perturbations in extra-cellular iron concentrations [17].

2.4 Resources for Transcriptional Regulatory Interactions in Prokaryotes

Recent progress in sequencing and expression profiling has delivered us a wealth of data and fueled the development of bioinformatics tools and online resource for the study of transcriptional regulation in prokaryotes. Unfortunately, a persistent challenge is that most studies and resources are biased towards a very small number of model organisms and thus may be of limited utility for the study of the great majority of prokaryotes. However, it can be expected that prokaryotes share at least a rudimentary basic regulatory network, as they appeared to have derived from a common ancestor. From this common network, branches might have then evolved to adapt to specific external conditions [18]. Thus, genome level comparison of well studied model organisms, such as *Escherichia coli* [19]; [20]; [21] and *Bacillus subtillis* [22-23], might still be useful for interpretation of regulatory networks in less studied organisms.

Table 1 gives a short list of selected tools and resources for the study of transcriptional regulation in prokaryotes. It should be noted that this research field is highly dynamic; and new tools and databases are rapidly developed. Thus, table 1 should not be considered as a comprehensive overview, but rather collection of initial pointers for the interested reader.

[TABLE 1]

2.5 Inferring Transcriptional Regulatory Networks

Capturing the transcriptional regulatory network of a given organisms should equip us with an explicit and comprehensive model of its transcriptional regulation. In the ideal case, this network would contain all regulatory components (such as TFs, ncRNAs and their binding sites) as well as all the regulatory interactions between them and with their targets. Such a model could enable us to predict accurately transcript levels of target genes. Clearly, comprehensive regulatory networks would not only be valuable for microbiology, but could be of crucial importance for future bioengineering and synthetic biology permitting accurate prediction of cellular responses to genetic manipulations. At present, however, we are still far from such models. Despite all the recent technological breakthroughs, deriving models for transcriptional regulatory networks remains a formidable task even for simple processes in well studied organism.

Commonly, transcriptional regulatory networks are represented as directed graphs comprised of nodes and edges. Nodes can represent transcription factors or ncRNAs and their target genes, while edges indicate regulatory interactions between the different components of the network [24]. To re-construct these networks, most approaches can be broadly classified into three types:

(i) *Knowledge-based approaches:* Traditionally, models of transcriptional regulatory networks have been derived using the knowledge accumulated in scientific literature and databases. Here, an initial network model is constructed to simulate the system´s behaviour under different perturbations and is subsequently adjusted to consolidate it with experimental observations. Examples for this approach can be found further below, where we discuss differential equations for modeling.

(ii) *Reverse engineering:* Recently, reverse engineering has become an promising alternative given the rapid increase of available expression data. Here, regulatory interactions are inferred directly from observed expression patterns of genes using correlation measures. The predicted transcriptional response is then compared with measured expression data and the network structure is iteratively improved. Reverse engineering methods might reduce the time needed to obtain accurate regulatory network models, but demand also large amount of data for inference [25]. Network inference can be direct or module-based. Examples are given in the next section.

(iii) *Template-based methods*: Thanks to astonishing advances in sequencing technology, the number of available prokaryotic sequenced genomes has greatly increased over recent years. Notably, many of these newly sequenced genomes belong to organisms, for which only limited knowledge of their biology exist. In such cases, transcriptional regulatory networks of a phylogenetically related organism might be used as a template for inference. Such methods are based on evolutionary conservation of regulation and can take advantage of abundant information on regulatory interaction from model organisms. A good example of this approach

is the work by Babu and coworkers who inferred conserved regulatory networks of an organism by comparing its genes to known TFs and their gene targets in *E. coli* [26]. Homologs of *E. coli* genes were identified by sequence comparison. If TF and its target are found to be conserved, their regulatory interaction is predicted to be also conserved. Clearly, template-based methods lose accuracy when applied to phylogenetically distant organisms. Here, additional analysis of potential binding sites for transcription factors might improve prediction of target genes [27].

2.5.1 Direct and Module-based Network Inference

Direct and module-based methods use gene expression to detect and model the underlying regulatory interactions. Both methods assume that changes in gene expression are caused by changes in the expression of TFs. A direct inference approach was applied for *E. coli* by Faith and co-workers [47]. Initially, 179 microarrays measurement for 69 different conditions from nine previous studies were collected and expanded by further in-house 266 microarrays measurements for 121 selected conditions. The applied context likelihood of relatedness (CLR) algorithm is based on previously developed concept of relevance networks [48-49]. The CLR method assumes that expression of a gene and of its potential regulator should vary in a coordinated manner over time and across environmental conditions. As measure of correlation of expression, CLR calculates the mutual information between expression profiles. To assess significance of co-expression, the observed mutual information is compared to a background distribution of mutual information scores for all possible TF-gene pairs. Only pairs with mutual information scores that are significantly higher than the background distribution were subsequently included in the transcriptional regulatory network. The authors compared the detected interactions by CLR algorithm with those annotated in the curated RegulonDB database and found that the filtering for significant interactions reduced considerably the number of false positives. At a 60% true positive rate, CLR predicted 1079 regulatory interactions of which 741 were novel.

An alternative approach to capture relationships between regulators and their target genes is based on clustering of gene expression data. Clustering is commonly used to group genes into sets with similar expression patterns. Such sets of co-expressed genes, which are also called modules, may underlie the same transcriptional regulation and are the basic units of module networks. For these types of networks, we infer regulatory interactions between TF and whole modules instead of interactions between TF and their individual target genes. The usage of modules instead of single genes reduces drastically the number of variables in the model, and thus, the computational burden as well as the risk of over-fitting.

This approach was used by Bonneau and colleagues to study the transcriptional regulatory network of the poorly characterized *Halobacterium salinarum NRC-1* [50]. Genome-wide expression data for a variety of different environmental conditions and genetic mutations were clustered using the cMonkey

algorithm [51]. This algorithm performs parallel "biclustering" of genes and conditions. The derived clusters (or modules) of genes can therefore be specific to conditions, in which the genes are co-expressed. This capacity of biclustering can allow subsequent identification of conditions, in which a regulator is active. The final transcriptional regulatory network was derived by the "Inferelator" method which connects expression changes of individual or multiple TFs to expression changes in the detected modules using a regression algorithm [52].

2.6 Modeling of Transcriptional Regulatory Networks

Once the structure of a regulatory network has been inferred, its behavior can be modeled. A key step is the selection of the mathematical approach to be employed. This decision will mainly depend on the amount, type and quality of available data and information. Several alternatives have been developed based on different mathematical frameworks such as differential equations [28], Boolean networks [29-30] or Bayesian networks [31-33]. In the following section, we will briefly illustrate their use in modeling transcriptional networks.

2.6.1 Differential Equations

Usage of differential equations to describe regulatory relationships has originated from the kinetic modeling of biochemical reactions [34]. A transcriptional regulatory network can be modeled by a system of linear differential equations with its rate parameters known or estimated. As exact parameters have typically to be estimated, it is important to limit the number of regulatory interactions. Therefore, a common assumption is that gene expression is determined by small set of global transcriptional regulators. Also, the use of differential equations for modeling is computationally expensive. Thus, they are commonly used only for modeling of small systems with a limited number of components and interactions.

Ropers and co-workers used piecewise-linear differential equation to simulate the adaption of *E.coli* during transition from a carbon-rich to a carbon-poor environment [35]. Their model consisted of six piecewise-linear differential equations which represented the expression of key global regulators over time during the response to carbon limitation. The use of piecewise-linear equations permitted qualitative analysis of the network dynamics even as quantitative information was scarce. For simulation, the open source software tool "Genetic Network Analyzer" was used [36]. The resulting network has been extended in a later study to include also directionality of regulation and influence of metabolites on the network flux [37].

In a recent study, differential equations were used to model cell cycle regulation in *Caulobacter crescentus* [38]. The model described the dynamics of three global regulators (GcrA, DnaA, and CtrA) and other cell cycle proteins by sixteen nonlinear ordinary differential equations. Although the model

is of relative small scale, more than 40 parameters (rate constants, binding constants and thresholds) had to be optimized, after initial values were estimated based on experimental observations from the literature. As an application of the trained model, the phenotypes of novel mutants could be successfully predicted. This shows that modeling of differential equation modeling is feasible for cases where sufficient information is available and the size of the system is small.

2.6.2 Boolean Networks

In the framework of Boolean networks, gene expression levels are binned to the values 1 or 0. Therefore, Boolean networks assume that a gene can be either expressed, i.e. in the "on" state described with the value 1, or not expressed, i.e. in the "off" state described with the value 0 [39]. During simulation, the level of a gene is derived from the levels of other genes via a Boolean function. In particular, expression of a gene at a given time point depends on the expression of its regulators at the previous time point. The "on-off" assumption limits the capabilities of Boolean networks. Also, they fail to model transcription factors that regulate their own expression. Nevertheless, in cases where only qualitative knowledge is available and the size of the network is of moderate size, Boolean networks can explore the full state space.

The work by Samal and Jain provide a good example of how Boolean networks can assist in system level analysis and modeling [40]. Their study is based on a previously compiled data set for *E. coli* [41]. The constructed Boolean network comprised almost 600 genes and 100 metabolites. Studying the dynamical properties of the network, they found that the network states are highly robust to perturbation of single genes, while the system is still highly responsive to environmental changes.

2.6.3 Probabilistic Boolean Networks

Frequently, the lack of experimental information on certain regulatory entities makes it necessary to include probabilistic elements. Probabilistic elements were introduced to Boolean networks to overcome their deterministic nature [42]. Each entity is modeled by a family of Boolean functions, to which probabilities are assigned. The function is chosen with assigned probability to predict the particular expression value of the target gene at certain time point or condition. The stochastic model allows a greater range of possible solutions and can cope with uncertainty.

Chandrasekaran and Price integrated metabolic and transcriptional regulatory networks using an approach closely related to probabilistic Boolean networks [43]. Their method termed PROM (Probabilistic Regulation Of Metabolism) circumvents the tedious process involved in the determination of the Boolean rules between the regulator and its targets by deriving automatically interactions from high-throughput data. For *E. coli*, PROM was tested and compared with other state of the art methods that combine metabolic and gene regulatory data using Boolean rules. The main

advantage of PROM resides in its automatism to calculate conditional probabilities between gene states from expression data. The authors also demonstrated the capacity of their approach by genome-scale modeling the metabolic and regulatory network of the *Mycobacterium tuberculosis* from available regulatory information and microarray data. Their model permitted the prediction of phenotypes for several TF knockout mutants as well as the identification of genes candidates for drug targeting.

2.6.4 Bayesian Networks

Bayesian networks display conditional dependencies between variables represented as nodes. Each node is associated with a function determining the probability of the corresponding variable. A drawback of Bayesian networks is the failure to model regulatory feedback loops, since they can infer only directed acyclic graphs. This limitation is partially solved by the use of Dynamical Bayesian networks (DBNs) [44].

Hodges *et al*. [45] used an implementation of the Bayesian network to analyze microarray gene expression data from *E. coli* and model the response to 27 genes linked to the reactive oxygen species (ROS) detoxification pathway defined in the EcoCyc database [46]. Expression data from over 300 measurements were utilized to score randomly initiated Bayesian networks. Subsequently, a consensus network was constructed from the networks with the highest posterior probability score. Even though the interactions of the consensus network did not fully match the known ROS pathway, it served as base for expansion. Some predictions of novel genes and interactions in the ROS pathway were successfully experimentally validated.

3 Protein Interaction Networks in Human

Many - if not most – proteins do not function alone, but through interaction with other proteins in a defined physiological context. In fact, extensive experimental and computational analysis of protein-protein interactions (PPIs) in *S. cerevisae* showed that a large part of the proteome is organized in cluster structures (or so-called modules), in which proteins are highly connected [53]. The study of PPIs is therefore an essential prerequisite for understanding many cellular processes. Here, we present a brief review of the experimental detection and computational prediction of PPIs as well as examples of relevant databases and network-orientated applications of PPI data. A graphical overview is presented in figure 2.

[FIGURE 2]

3.1 Types of protein-protein interactions and their detection

PPIs can be broadly classified based on the complexes formed and the duration of interaction: (i) Interaction of identical polypeptide chains form homo-oligomers, whereas interactions of non-identical chains form hetero-oligomers. (ii) Permanent interactions persist until the complex is degraded, whereas transient interactions can form and break *in vivo* [54]. For detecting different types of interactions, various experimental techniques have been developed. It is important to appreciate their characteristics as well as strength and weakness, as different techniques can produce distinct sets of interactions even for the same set of proteins. In the following section, we introduce the yeast-two-hybrid (Y2H) and co-immunoprecipitation (Co-IP) approach, which have been used to screen for protein interactions in a high-throughput mode.

3.1.1 Yeast-Two-hybrid (Y2H) assay

Y2H is a well established system, which has initially been introduced in 1989 [55] to identify interactions between two selected proteins and has been now scaled up to cover the entire proteome of an organism. In Y2H system, one protein (the so-called bait) is fused with the DNA binding domain of a yeast transcription factor, while the second protein (prey) is fused with its activation domain. Physical binding of bait and prey proteins re-constitutes a functional transcription factor whose activity is monitored through a reporter gene. In the last decade, the technique has been successfully employed in large-scale mapping of PPIs for model organisms such as *S. cerevisiae* [56], *C. elegans* [57] and *D. melanogaster* [58] as well as for human [59-60]. In a high-throughput mode, large sets of yeast strains with bait and prey proteins are constructed. A common approach is to mate individual

strains containing bait proteins with a pooled library of yeast strains containing prey proteins [61]. For positive read-outs, the interacting prey protein will be determined by sequencing. Besides its usability for large screens, an attractive feature of Y2H assay is that weak transient interactions can be detected. However, false results may occur when additional proteins present in yeast positively or negatively affect binding of proteins assayed, or if bait or prey proteins itself have transcriptional activator or repressor activity. Also, interacting proteins have to be located to the nucleus for detection which can cause difficulties in screening of membrane proteins.

Notably, the Y2H system was applied in two large-scale studies to screen human proteins [59-60]. Stelzl and co-workers used a combination of fetal brain cDNA library and a set of full length open reading frames (ORFs) to create over 11 000 Y2H clones [59]. Applying a pooling approach, more than 25 million protein pairs were tested resulting in the identification of over 3000 interactions. Independently, Rual and collaborators performed an Y2H screen with more than 8000 ORFs and detected ~2800 interactions [60]. However, caution needs to be taken in the interpretation of the reported interactions of human proteins, as they were measured in yeast (i.e. outside their native surroundings) and posttranscriptional modification altering binding might not be conserved in yeast.

3.1.2 Co-immunoprecipitation (Co-IP)

Co-immunoprecipitation is a commonly used method to test whether proteins are bound in the same complex *in vivo*. The target protein is immunoprecipitated with an antibody and fractionated by SDS-PAGE. Co-immunoprecipitated proteins are subsequently detected by autoradiography or Western blotting. Also, protein sequencing may be used to identify the interacting proteins. It is important to note that this technique assumes that interactions are preserved when a cell is lysed under non-denaturing conditions. To identify protein interaction in a high throughput mode, affinity purification is coupled with subsequent mass spectrometry [62-63]. On a system-wide level, this combination was first exercised for yeast [53]. A general observation was that a considerable part of the proteome can be organized in protein complexes. For instance, Gavin and colleagues could identify over 200 mostly novel complexes for an initial set of more than 1.700 tagged yeast proteins. Detailed analyses revealed that most protein complexes form a higher order network beyond the level of binary interactions. To inspect human complexes, Ewing chose an initial set of 338 bait proteins expressed in a human cell line (HEK293) [64]. Despite the rather small number of bait proteins, almost 25.000 interactions were detected. After filtering, ~6.500 interactions between over 2.200 proteins remained, most of which were newly identified.

Note that identified interactions are not necessarily direct interactions, but might be indirect, interactions between proteins in the precipitated complex. This has consequences if we want to represent the complex by binary interactions in order to facilitate network analysis. As the internal structure of the complex is generally not known, two generic models are commonly employed: The

matrix model postulates that all proteins in a complex interact with each other. Naturally, this postulation results in many interactions especially for large complexes and thus implies potentially a large number of false positive binary interactions [65]. In contrast, the *spike* model assumes that direct interactions exist only between the bait and the co-precipitated proteins neglecting all other internal structures in the complex.

3.2 Computational prediction of PPI

Despite rapid advances in large-scale experimental techniques applied in mapping the human PPI network, the coverage of experimentally determined PPI data remains scarce compared to estimated total number of approximately 650.000 interactions [66]. However, computational methods might assist in the mapping of PPI networks. Computational approaches not only enable the discovery of novel putative interactions, but can also provide information for designing experiments for specific proteins. They are either based on predictions using existing data or on text mining of published literature.

For the human interactions, the most important method is based on sequence conservation between organisms. This approach assumes that interactions are evolutionary conserved between orthologous proteins in different organisms. Initially, the concept of so-called 'interologs' was introduced to examine the biological relevance of Y2H-derived interactions [67]. Although subsequent experimental validation only resulted in a limited accuracy (up to 30%) for predicted PPIs in *C. elegans*, several large sets of human PPIs has been computationally extrapolated from experimentally measured interactions in lower organisms, especially yeast, worm, fly and mouse [68-70]. For the identification of human orthologs for interacting proteins, the InParanoid algorithm has commonly been employed [69-71]. In a first application of the interolog concept to human PPIs, interactions from three model organisms (*S. cerevisiae*, *D. melanogaster* and *C. elegans*) were utilized [69]. An interaction was predicted if both interacting proteins in a model organism have one or more human orthologs. Using this strategy, the authors generated a human interaction network comprising ~71,000 interactions between ~6,000 human proteins. The generated map was then filtered using Gene Ontology annotation and co-expression in order to identify a core network of over 9500 interactions between ~3500 unique proteins. A similar study was undertaken by Perisco *et al.* and lead to the construction of the HomoMINT database [70]. Besides using interactions from lower organisms, they analysed the domain composition of human proteins to refine the predictions of interaction. Instead of the InParanoid alogrithm, Brown and Jurisica applied a BLAST and reciprocal best-hit approach to extrapolate interactions between organisms [72]. They created first an integrated interaction dataset from various model organisms and mapped it to human orthologs by aligning proteins from each model organism against human proteins stored in SWISS-Prot database. As a next step, each top BLAST hit surpassing a pre-defined significance threshold was matched against the set of all protein

sequences of the model organism. A protein was considered as potential ortholog, if it matched the original query protein in reverse direction. Following this method, the autors generated a human PPI map containing ~25000 interactions between ~4000 proteins.

Ideally, we would like to accurately predict potential protein interactions directly from the primary sequences. Notably, several groups have undertaken this challenging task and have reported the successful prediction of PPIs based on sequence data only [73-77]. The different approaches commonly represent a protein sequence as a vector of features (such as the physicochemical properties of amino acids) and a protein pair as concatenation of the corresponding feature vectors. For subsequent classification and prediction, support vector machine or other kernel based method can be trained to distinguish between concatenated feature vectors of interacting and non-interacting protein pairs. The reported high accuracy, however, is somewhat surprising given that many PPI depend on mutual recognition of detailed binding interfaces. Indeed, follow-up studies have shown that the reported performance depends strongly on the selection of training and test sets [77-78]. Thus, the results need to be interpreted with caution. Nevertheless, the proposed prediction methods might help to derive an initial set of potential PPIs for experimental validation.

In addition to prediction based on existing data, potential PPIs can be computationally indicated evaluating existing literature. A simple text-mining approach is the measurement of the frequency of protein names co-occurring in the same scientific text. If the frequency is higher than expected by chance, a functional association and potential physical interaction might be implicated. An early application of this approach generated a network of over 6600 interactions connecting ~3700 human proteins based on text-mining of Medline abstracts [79]. However, one needs keep in mind that the deduced interactions need not to be physical. The advantages of such data mining techniques are that they are not biased to any particular study or experimental technique. Major drawback can be significant selection and detection bias towards well studied proteins [80].

3.3 Literature curation and PPI databases

Besides experimental high-thoughput approaches and computational prediction, the manual curation of literature is one main strategy to obtain large sets of human PPIs. Indeed, many interactions have been derived in dedicated small-scale experiments and have been reported in the scientific literature. To curate this information in a systematic way, several databases have been established including the Human Protein Reference Database (HPRD)[81], the Biological General Repository for Interaction Datasets (BioGRID)[82], IntAct [83], Database of Interaction Proteins (DIP)[84], the Biomolecular Interaction Network Database (BIND) [85] and the Mammalian Protein-Protein Interaction Database (MPPI) [86].

Notably, these databases are not synchronized with each other, and their data formats might also be incompatible. To merge publicly available protein interaction data, a time-consuming reformatting and mapping must be undertaken. To improve data exchange, Hermjakob *et al.* [87] proposed the Molecular Interaction (MI) extensible markup language (XML) format as a standard for the representation and exchange of protein interaction data. Notably, the PSI MI format is a database-independent. This format can also help combine data from different sources.

To relieve users of cumbersome data pre-processing and merging, several databases have integrated both experimental as well as predicted PPIs sets. Examples are I2D [68], STRING [88] and UniHI [89]. For instance, the Unified Human Interactome database (UniHI) includes human interaction data from 14 different sources and comprises approximately 250,000 interactions between over 22,000 proteins [89]. The latest version of UniHI not only integrates the data from various sources, but also provides easy query options in order to extract a maximum of information. Furthermore, users can also map their own gene expression data onto the PPI network. Additionally, analyses can be performed to detect associations of a PPI network with biological processes, molecular functions and cellular components. To illustrate the application of UniHI, we used expression data for a mutant p53 cell line [90]. The original study focused on the role of mutant p53 in promoting transformation and metastasis in breast cancer through aberrant protein interactions. UniHI enables to derive a PPI network around p53 and to readily map the gene expression data onto the network. After filter of interactions by differential expression, we obtained a specific dys-regulated p53-focused network. Using the tool for functional analysis in UniHI, we detected DNA damage response, cell cycle and regulation of apoptosis as processes that are significantly enriched in the network (figure 3). As this example illustrates, we can rapidly indentify p53 interactions and functions that are potentially affected by dys-regulation and that can serve as starting point for further study.

[FIGURE 3]

3.4 Network-based study of cancer

For the study of pathogenic processes, the usage of PPI networks has become an attractive paradigm. Diseases occur frequently not due to malfunctioning of a single protein, but of whole complexes or modules. An example showing the importance of such complexes is the Faconi anaemia, a genetic disease associated with increased rate of leukaemia and bone marrow failure. Remarkably, the corresponding proteins of seven of the nine genes associated with Faconi anaemia form a physical complex involved in DNA repair. In fact, cancer *per se* can been seen as a classical example how alterations in several genes and proteins need to occur for the disease to develop and progress. Genes involved in cancer can be generally assigned into three broad classes based on their functions: i) *oncogenes* that can lead in their mutated form to aberrant growth signals; ii) *tumour-suppressor genes*

that deter cells from unrestricted proliferation; and iii) *stability genes* that keep genetic alterations to a minimum. It is important to note, that no single gene defect causes cancer. Rather, multiple mutations in oncogenes, tumour-suppressor as well as stability genes have to occur to overcome the control mechanisms imbedded in cells [91-92]. In support of this concept, recent sequencing of a large number of genes in breast and colorectal tumours revealed various mutations accumulated in a single tumour [93]. Eventually, the changes on molecular level lead to a series of complex alterations in cell physiology ranging from self-sufficiency in growth-signals and evasion of apoptosis to tissue invasion and metastasis [91]. Again, each of these steps displays a high variability and is driven by multiple molecular mechanisms. To study this complexity and to find new targets in cancer treatment, network-based approaches promise to provide us with a powerful tool.

3.4.1 Identification of Novel Cancer-Associated Genes

PPI data have been utilized in various studies to identify novel genes associated which cancer. Such genes can aid in the early diagnosis and in the development of effective therapies. For instance, Russo AL *et al*. combined human PPI with others types of data to study prostate cancer [94]. Using gene expression and PPI data as well as pathway information, they identified the Met receptor tyrosine kinase as a metastatic biomarker and central regulator of prostate cancer progression. Another example of network-based approaches is the detection of GTP binding protein 4 (GTPBP4) as a marker for the breast cancer disease prognosis [95]. In this study, the authors first identified novel interactors of p53 in Drosophila and searched for corresponding orthologs in human. One of these novel interacting partners of p53 was GTPBP4. Knockdown of GTPBP4 in a cancer cell line led to activation of p53. Also, it was observed that expression levels of GTPBP4 were inversely correlated with survival of breast cancer patient. In the study of glioblastoma, a highly aggressive form of brain cancer, Ladha and co-workers used a set of up-regulated genes as seed proteins for network construction [96]. Analyses of the derived PPI network revealed that two proteins (casein kinase 2 catalyticsubunit α and protein phosphatase 1 alpha) connected to network structures potentially associated with cancer progression. Both proteins were shown to be up-regulated in a set of glioblastoma tumour samples. In a conceptually similar work, a serous ovarian cancer-related network was constructed [97]. Initially, 30 genes were obtained which shown consistent up- or down-regulation of expression in different data sets. They were used as initial nodes for network construction. Subsequent integration of gene expression data with PPI data identified highly connected proteins (hubs) involved in early events of cell cycle progression, mismatch repair and aneuploidy in ovarian cancer as well as novel cancer-related genes, whose precise involvement remains to be investigated.

As the studies show, the integration of PPI with complementary data is a common approach to detect cancer genes. A final example is a study of breast cancer, where the authors integrate PPI networks from different species with co-expression and genetic interactions as well as functional gene

annotation [98]. Starting from a set with four known breast cancer-associated genes (BRCA1, BRCA2, ATM, and CHEK2), the derived network was used for prediction of novel breast cancer susceptibility genes. For one predicted gene, HMMR encoding a centrosome subunit, a functional association with BRCA1 was experimentally demonstrated.

3.4.2 Network-Based Analysis of Cancer-Related Mechanisms

In addition to prediction of novel cancer-associated genes, PPI networks have also been utilized to untangle cancer features. In a recent study [99], the authors analysed features of prostate cancer features using PPI networks and molecular profiles in adjacent normal cells. First, specific set of genes were derived for normal cells, normal cells adjacent to tumour and tumour cells. Second, PPI networks for these set of genes were analysed. Notably, they identified three sub-networks consisting of pro-inflammatory cytokine, pro-metastatic chemokines and growth factors. Furthermore, genes encoding cytokines and growth factors were found to be expressed in adjacent normal epithelial cells (i.e. in the tumour microenvironment) suggesting that the molecular state of adjacent normal cells might have prognostic value.

An emerging application of molecular networks analysis is the use of PPI networks to improve disease classification. Chuang *et al.* employed a network-based classifier for the prognosis of breast cancer metastasis, which is the main cause of death among breast cancer patients [100]. Gene expression profiles of metastatic and non-metastatic patients were mapped onto a human PPI network and sub-networks, whose expression levels correlate with metastasis, were identified. Here, the expression level of a sub-network was defined as a function of expression levels of the included genes. Strikingly, the authors found that classification of metastatic versus non-metastatic tumour samples is more reproducible and more accurate, if it is based on sub-networks than on individual gene expression signatures only.

Apart from disease classification, PPI networks have been employed in predicting cancer outcome using their modular structure [101]. A modular architecture can be imposed on the human protein interactome through inter-modular hubs (having low correlation of expression with interaction partners) and intra-modular hubs (having high correlation of co-expression with interaction partners). Taylor *et al.* studied the protein domain numbers, domain sizes and linear motifs (i.e. post translational modification and short binding motifs) of inter-modular and intra-modular hubs in the human protein interactome across 79 human tissues [101]. They identified that number of domains is higher for inter-modular hubs, whereas domain sizes are larger in intra-modular hubs. Linear motifs are over-represented in inter-modular hubs. Further, they compared the modularity of the protein interactome in breast cancer patients with respect to disease outcome and found changes in modularity that were linked to the outcome of breast cancer. Such changes may provide a prognostic signature for breast cancer.

4 Current Challenges and Future Directions

Molecular interactions are crucial for the correct and efficient functioning of many cellular processes. Due to their biological relevance, the study of interaction networks has evoked increasing interest in their computational analysis. In parallel, systematic efforts and specialized databases have led to a dramatic growth in the number of chartered interaction data during recent years. This growth offers now an unprecedented wealth of interaction data to researchers for their investigations.

As a long term perspective, we can aim to identify the full repertoire of molecular interactions for all components of a cell or even of a whole organism. Knowing this 'interactome' will doubtlessly yield better insights into the molecular machinery of cells and disease mechanisms. Naturally, a combination of experimental and theoretical approaches is needed to accomplish such ambitious goal and to consolidate the obtained interactions with gene expression, functional annotation or other types of biological data.

At present, however, the available information about biological networks is still very limited and we have to work with crude approximations in network analyses. For instance, we would frequently prefer to utilize activity of proteins instead of transcript level for the modeling of molecular networks. For large networks, however, such analysis have remained beyond current state of the art, as necessary information of e.g. protein abundances as well as of post-translational modifications (PTMs) and their effects is required, that is, information which is generally not available yet on a genome-wide level.

Also, molecular networks are inherently complex and difficult to model accurately, because of their numerous components and multiple layers. For illustration: Analyses of the human protein-protein interactome are typically confronted with the large number of predicted protein-coding genes (~25000), unknown spatial and temporal localization of many proteins as well as a numerous PTMs that might alter protein-protein interactions. In particular, alterations in transient interactions can lead to various changes in network structures and thus contribute to a highly dynamic network. The stability of such interactions depends on cellular physiological conditions and environment. This feature of molecular interactions should be always taken into account in the analysis and interpretation of interaction networks, especially as most presentations (such as figure 1) give a static - and potentially misleading - picture.

The dynamics of interactions is frequently regulated by post-translational modifications of proteins. In general, the effects of PTMs are manifold: they can influence protein size, hydrophobicity, and other physical-chemical properties; they can enhance, change, or block specific protein activities; and they can target proteins to specific sub-cellular localization. Currently there are over 80.000 experimentally characterized PTMs across different organisms reported in the SWISS-PROT database. Despite of

their importance, however, PTMs are often not included in models of interaction networks due to missing data as well their unknown effects. In general, the full structure of global molecular networks is largely unknown, as they consist of many interwoven networks of different type such as regulatory transcriptional, metabolic or physical protein interaction networks.

Despite current limitations, recent studies have indicated that network-based approaches can substantially contribute to our knowledge how biological systems function. In fact, molecular networks might be a natural extension of pathways models, a concept which have been highly successful in biology and medical research. For instance, pathway models helped to organize the large number of cancer-relevant genes in functional coherent groups [92]. In cancer research, pathway models were motivated by the observation that mutations within the same pathway frequtenly produce the same effects. For example, the control of cell cycle can not only become defective by direct mutation of the retinoblastoma (Rb) gene but equally by other mutation in the associated Cdk/Rb/E2F pathway. The pathway concept allows here for consolidation of initially unrelated observations and can further reveal communalities of different types of cancers. Additionally, pathway-focused analyses have facilitated the interpretation of mutations in their functional context.

Despite the success of the pathway concept, it is important to note that pathways derive by no means naturally and that canonical pathway models seem frequently too simplistic. Notably, an inclusion of network-based concepts might contribute to our capability to capture molecular mechanisms. Indeed, it has pointed out by Friedman and Perrimon that the canonical view of signalling pathways as compartmentalized linear circuits is in stark contrast with the results from recent systematic screens [102]. A reason for the traditional simplistic picture of linear pathways might have be the dominance of developmental screens based on a qualitative readout. Such set-up may allow only the most important signalling proteins to be discovered. In contrast, large numbers of signalling modifiers have been identified when assays with sensitive quantitative readout were used. Thus, it can be argued that signalling transduction should be rather seen as quantitative information propagating through densely connected molecular interaction networks.

In future, we may therefore refer to specific networks (e.g. the ‘Wnt network’), similarly as we are now readily using the concept of pathways (e.g. the Wnt pathway). Clearly, much has still to be learned about the composition and regulation of molecular networks. However, we anticipate that such step will constitute a crucial advancement towards a true understanding of the complexity lying within cells and organisms.

References

1. Kitano, H., *Computational systems biology.* Nature, 2002. **420**(6912): p. 206-10.
2. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.
3. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
4. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.
5. Steuer, R.a.J., B. H, *Computational Models of Metabolism: Stability and Regulation in Metabolic Networks*, in *Advances in Chemical Physics*, S.A. Rice, Editor. 2008, John Wiley & Sons: Hoboken, NJ, USA.
6. Médigue, C., et al., *Evidence for horizontal gene transfer in Escherichia coli speciation.* Journal of Molecular Biology, 1991. **222**(4): p. 851-856.
7. Ermolaeva, M.D., et al., *Prediction of transcription terminators in bacterial genomes.* Journal of Molecular Biology, 2000. **301**(1): p. 27-33.
8. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
9. Iyer, V.R., et al., *Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.* Nature, 2001. **409**(6819): p. 533-8.
10. Simon, I., et al., *Serial regulation of transcriptional regulators in the yeast cell cycle.* Cell, 2001. **106**(6): p. 697-708.
11. Schmidt, D., et al., *Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding.* Science, 2010. **328**(5981): p. 1036-1040.
12. Kahramanoglou, C., et al., *Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli.* Nucleic Acids Research, 2011. **39**(6): p. 2073-2091.
13. Stougaard, P., S. Molin, and K. Nordstrom, *RNAs involved in copy-number control and incompatibility of plasmid R1.* Proc Natl Acad Sci U S A, 1981. **78**(10): p. 6008-12.
14. Brownlee, G.G., *Sequence of 6S RNA of E. coli.* Nat New Biol, 1971. **229**(5): p. 147-9.
15. Herbig, A. and K. Nieselt, *nocoRNAc: Characterization of non-coding RNAs in prokaryotes.* BMC Bioinformatics, 2011. **12**(1): p. 40.
16. Novais, R.C. and Y.R. Thorstenson, *The evolution of Pyrosequencing® for microbiology: From genes to genomes.* Journal of Microbiological Methods, 2011. **86**(1): p. 1-7.
17. Masse, E. and S. Gottesman, *A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli.* Proc Natl Acad Sci U S A, 2002. **99**(7): p. 4620-5.
18. Daubin, V., M. Gouy, and G. Perriere, *A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.* Genome Res, 2002. **12**(7): p. 1080-90.
19. Thieffry, D., et al., *From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli.* Bioessays, 1998. **20**(5): p. 433-40.
20. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.
21. Dobrin, R., et al., *Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network.* BMC Bioinformatics, 2004. **5**: p. 10.
22. Harwood, C.R. and I. Moszer, *From gene regulation to gene function: regulatory networks in bacillus subtilis.* Comp Funct Genomics, 2002. **3**(1): p. 37-41.
23. Sellerio, A.L., et al., *A comparative evolutionary study of transcription networks. The global role of feedback and hierachical structures.* Mol Biosyst, 2009. **5**(2): p. 170-9.
24. Kim, T.-M. and P.J. Park, *Advances in analysis of transcriptional regulatory networks.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2011. **3**(1): p. 21-35.
25. Gardner, T.S. and J.J. Faith, *Reverse-engineering transcription control networks.* Physics of Life Reviews, 2005. **2**(1): p. 65-88.
26. Babu, M.M., et al., *Structure and evolution of transcriptional regulatory networks.* Curr Opin Struct Biol, 2004. **14**(3): p. 283-91.
27. Tavazoie, S., et al., *Systematic determination of genetic network architecture.* Nat Genet, 1999. **22**(3): p. 281-285.
28. Chen, T., H. He, and G. Church, *Modeling gene expression with differential equations.* Pacific Symposium Biocomputing, 1999. **4**: p. 29 - 40.
29. Kauffman, S.A., *Metabolic stability and epigenesis in randomly constructed genetic nets.* Journal of Theoretical Biology, 1969. **22**(3): p. 437-467.

30. Akutsu, T., S. Miyano, and S. Kuhara, *Identification of genetic networks from a small number of gene expression patterns under the Boolean network model.* Pacific Symposium on Biocomputing, 1999. **4**: p. 17 - 28.
31. Liang, S., S. Fuhrman, and R. Somogyi, *REVEAL, A general reverse engineering algorithm for inference of genetic network architectures.* Pacific Symposium on Biocomputing, 1998. **3**: p. 18 - 29.
32. Friedman, N., M. Goldszmidt, and A. Wyner, *Data analysis with Bayesian networks: A bootstrap approach.* Proc Fifteenth Conf on Uncertainty in Artificial Intelligence (UAI), 1999.
33. Imoto, S., T. Goto, and S. Miyano, *Estimation of Genetic Networks and Functional Structures Between Genes by Using Bayesian Networks and Nonparametric Regression.* Pacific Symposium on Biocomputing, 2002. **7**: p. 175 - 186.
34. Glass, L. and S.A. Kauffman, *The logical analysis of continuous, non-linear biochemical control networks.* J Theor Biol, 1973. **39**(1): p. 103-29.
35. Ropers, D., et al., *Qualitative simulation of the carbon starvation response in Escherichia coli.* Biosystems, 2006. **84**(2): p. 124-152.
36. de Jong, H., et al., *Genetic Network Analyzer: qualitative simulation of genetic regulatory networks.* Bioinformatics, 2003. **19**(3): p. 336-44.
37. Baldazzi, V., et al., *The Carbon Assimilation Network in Escherichia coli Is Densely Connected and Largely Sign-Determined by Directions of Metabolic Fluxes.* PLoS Comput Biol, 2010. **6**(6): p. e1000812.
38. Li, S., et al., *A quantitative study of the division cycle of Caulobacter crescentus stalked cells.* PLoS Comput Biol, 2008. **4**(1): p. e9.
39. Kauffman, S.A., *The Origins of Order: Self-Organization and Selection in Evolution.* Biophys J, 1993. **65**(6): p. 2.
40. Samal, A. and S. Jain, *The regulatory network of E. coli metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response.* BMC Systems Biology, 2008. **2**(1): p. 21.
41. Covert, M.W., et al., *Integrating high-throughput and computational data elucidates bacterial networks.* Nature, 2004. **429**(6987): p. 92-6.
42. Shmulevich, I., et al., *Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks.* Bioinformatics, 2002. **18**(2): p. 261-74.
43. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proc Natl Acad Sci U S A, 2010. **107**(41): p. 17845-50.
44. Ong, I.M., J.D. Glasner, and D. Page, *Modelling regulatory pathways in E. coli from time series expression profiles.* Bioinformatics, 2002. **18**(suppl 1): p. S241-S248.
45. Hodges, A.P., et al., *Bayesian network expansion identifies new ROS and biofilm regulators.* PLoS One, 2010. **5**(3): p. e9513.
46. Keseler, I.M., et al., *EcoCyc: a comprehensive database of Escherichia coli biology.* Nucleic Acids Research, 2011. **39**(suppl 1): p. D583-D590.
47. Faith, J.J., et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8.
48. Butte, A. and I. Kohane, *Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput 2000: p. 11.
49. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.
50. Bonneau, R., et al., *A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell.* Cell, 2007. **131**(7): p. 1354-1365.
51. Reiss, D.J., N.S. Baliga, and R. Bonneau, *Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks.* BMC Bioinformatics, 2006. **7**: p. 280.
52. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.* Genome Biol, 2006. **7**(5): p. R36.
53. Gavin, A.C., et al., *Proteome survey reveals modularity of the yeast cell machinery.* Nature, 2006. **440**(7084): p. 631-6.
54. Nooren, I.M. and J.M. Thornton, *Diversity of protein-protein interactions.* EMBO J, 2003. **22**(14): p. 3486-92.
55. Fields, S. and O. Song, *A novel genetic system to detect protein-protein interactions.* Nature, 1989. **340**(6230): p. 245-6.
56. Yu, H., et al., *High-quality binary protein interaction map of the yeast interactome network.* Science, 2008. **322**(5898): p. 104-10.
57. Li, S., et al., *A map of the interactome network of the metazoan C. elegans.* Science, 2004. **303**(5657): p. 540-3.

58. Giot, L., et al., *A protein interaction map of Drosophila melanogaster.* Science, 2003. **302**(5651): p. 1727-36.
59. Stelzl, U., et al., *A human protein-protein interaction network: a resource for annotating the proteome.* Cell, 2005. **122**(6): p. 957-68.
60. Rual, J.F., et al., *Towards a proteome-scale map of the human protein-protein interaction network.* Nature, 2005. **437**(7062): p. 1173-8.
61. Walhout, A.J. and M. Vidal, *High-throughput yeast two-hybrid assays for large-scale protein interaction mapping.* Methods, 2001. **24**(3): p. 297-306.
62. Figeys, D., L.D. McBroom, and M.F. Moran, *Mass spectrometry for the study of protein-protein interactions.* Methods, 2001. **24**(3): p. 230-9.
63. Kocher, T. and G. Superti-Furga, *Mass spectrometry-based functional proteomics: from molecular machines to protein networks.* Nat Methods, 2007. **4**(10): p. 807-15.
64. Ewing, R.M., et al., *Large-scale mapping of human protein-protein interactions by mass spectrometry.* Mol Syst Biol, 2007. **3**: p. 89.
65. Bader, G.D. and C.W. Hogue, *Analyzing yeast protein-protein interaction data obtained from different sources.* Nat Biotechnol, 2002. **20**(10): p. 991-7.
66. Stumpf, M.P., et al., *Estimating the size of the human interactome.* Proc Natl Acad Sci U S A, 2008. **105**(19): p. 6959-64.
67. Walhout, A.J., et al., *Protein interaction mapping in C. elegans using proteins involved in vulval development.* Science, 2000. **287**(5450): p. 116-22.
68. Brown, K.R. and I. Jurisica, *Unequal evolutionary conservation of human protein interactions in interologous networks.* Genome Biol, 2007. **8**(5): p. R95.
69. Lehner, B. and A.G. Fraser, *A first-draft human protein-interaction map.* Genome Biol, 2004. **5**(9): p. R63.
70. Persico, M., et al., *HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms.* BMC Bioinformatics, 2005. **6 Suppl 4**: p. S21.
71. O'Brien, S.E., et al., *Computational tools for the analysis and visualization of multiple protein-ligand complexes.* J Mol Graph Model, 2005. **24**(3): p. 186-94.
72. Brown, K.R. and I. Jurisica, *Online predicted human interaction database.* Bioinformatics, 2005. **21**(9): p. 2076-82.
73. Martin, S., D. Roe, and J.L. Faulon, *Predicting protein-protein interactions using signature products.* Bioinformatics, 2005. **21**(2): p. 218-26.
74. Han, D.S., et al., *PreSPI: a domain combination based prediction system for protein-protein interaction.* Nucleic Acids Res, 2004. **32**(21): p. 6312-20.
75. Kanaan, S.P., et al., *Inferring protein-protein interactions from multiple protein domain combinations.* Methods Mol Biol, 2009. **541**: p. 43-59.
76. Guo, Y., et al., *Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences.* Nucleic Acids Res, 2008. **36**(9): p. 3025-30.
77. Yu, C.Y., L.C. Chou, and D.T. Chang, *Predicting protein-protein interactions in unbalanced data using the primary structure of proteins.* BMC Bioinformatics, 2010. **11**: p. 167.
78. Park, Y. and E.M. Marcotte, *Revisiting the negative example sampling problem for predicting protein-protein interactions.* Bioinformatics, 2011. **27**(21): p. 3024-8.
79. Ramani, A.K., et al., *Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome.* Genome Biol, 2005. **6**(5): p. R40.
80. Futschik, M.E., G. Chaurasia, and H. Herzel, *Comparison of human protein-protein interaction maps.* Bioinformatics, 2007. **23**(5): p. 605-11.
81. Keshava Prasad, T.S., et al., *Human Protein Reference Database--2009 update.* Nucleic Acids Res, 2009. **37**(Database issue): p. D767-72.
82. Stark, C., et al., *The BioGRID Interaction Database: 2011 update.* Nucleic Acids Res, 2011. **39**(Database issue): p. D698-704.
83. Aranda, B., et al., *The IntAct molecular interaction database in 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D525-31.
84. Salwinski, L., et al., *The Database of Interacting Proteins: 2004 update.* Nucleic Acids Res, 2004. **32**(Database issue): p. D449-51.
85. Isserlin, R., R.A. El-Badrawi, and G.D. Bader, *The Biomolecular Interaction Network Database in PSI-MI 2.5.* Database (Oxford), 2011. **2011**: p. baq037.
86. Pagel, P., et al., *The MIPS mammalian protein-protein interaction database.* Bioinformatics, 2005. **21**(6): p. 832-4.
87. Hermjakob, H., et al., *The HUPO PSI's molecular interaction format--a community standard for the representation of protein interaction data.* Nat Biotechnol, 2004. **22**(2): p. 177-83.

88. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
89. Chaurasia, G., et al., *UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome.* Nucleic Acids Res, 2009. **37**(Database issue): p. D657-60.
90. Girardini, J.E., et al., *A Pin1/mutant p53 axis promotes aggressiveness in breast cancer.* Cancer Cell, 2011. **20**(1): p. 79-91.
91. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
92. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control.* Nat Med, 2004. **10**(8): p. 789-99.
93. Wood, L.D., et al., *The genomic landscapes of human breast and colorectal cancers.* Science, 2007. **318**(5853): p. 1108-13.
94. Russo, A.L., et al., *Urine analysis and protein networking identify met as a marker of metastatic prostate cancer.* Clin Cancer Res, 2009. **15**(13): p. 4292-8.
95. Lunardi, A., et al., *A genome-scale protein interaction profile of Drosophila p53 uncovers additional nodes of the human p53 network.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6322-7.
96. Ladha, J., et al., *Glioblastoma-specific protein interaction network identifies PP1A and CSK21 as connecting molecules between cell cycle-associated genes.* Cancer Res, 2010. **70**(16): p. 6437-47.
97. Bapat, S.A., et al., *Gene expression: protein interaction systems network modeling identifies transformation-associated molecules and pathways in ovarian cancer.* Cancer Res, 2010. **70**(12): p. 4809-19.
98. Pujana, M.A., et al., *Network modeling links breast cancer susceptibility and centrosome dysfunction.* Nat Genet, 2007. **39**(11): p. 1338-49.
99. Trevino, V., et al., *Analysis of normal-tumour tissue interaction in tumours: prediction of prostate cancer features from the molecular profile of adjacent normal cells.* PLoS One, 2011. **6**(3): p. e16492.
100. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.
101. Taylor, I.W., et al., *Dynamic modularity in protein interaction networks predicts breast cancer outcome.* Nat Biotechnol, 2009. **27**(2): p. 199-204.
102. Friedman, A. and N. Perrimon, *Genetic screening for signal transduction in the era of network biology.* Cell, 2007. **128**(2): p. 225-31.

Resource	URL	Description
TFBS/motif discovery tools		
AlignACE	http://atlas.med.harvard.edu/	Identification of DNA motifs
MEME	http://meme.sdsc.edu/	Identification of DNA motifs
Promoter prediction		
PPP	http://bioinformatics.biol.rug.nl/websoftware/ppp/	Prediction of promoter
PromEC	http://margalit.huji.ac.il/	Database of E. coli mRNA promoters
MOTIFATOR	http://www.motifator.nl	Mining and characterization of regulatory DNA motifs
ncRNAs prediction		
nocoRNAc	http://it.inf.uni-tuebingen.de/de/nocornac.php	Prediction and characterization of ncRNA
Databases		
MicrobesOnline	http://www.microbesonline.org/	Compedium of genomes, expression data, operon predictions
RegPrecise	http://regprecise.lbl.gov/RegPrecise/index.jsp	Database of curated Regulons in Prokaryotic genomes
RegTransBase	http://regtransbase.lbl.gov	Curated database of regulatory interactions and database of TFBS
RegulonDB	http://regulondb.ccg.unam.mx/	E. Coli transcriptional regulatory network
Network analysis		
Bioconductor	http://www.bioconductor.org/	Network visualization and analysis
Cytoscape	http://www.cytoscape.org/	Network visualization and analysis
Pajek	http://pajek.imfm.si/	Network visualization ananalysis

Table 1. Bioinformatic tools and on-line resources for the discovery, analysis, or visualization of transcription factor binding sites, regulatory motifs, ncRNAs and transcriptional networks.

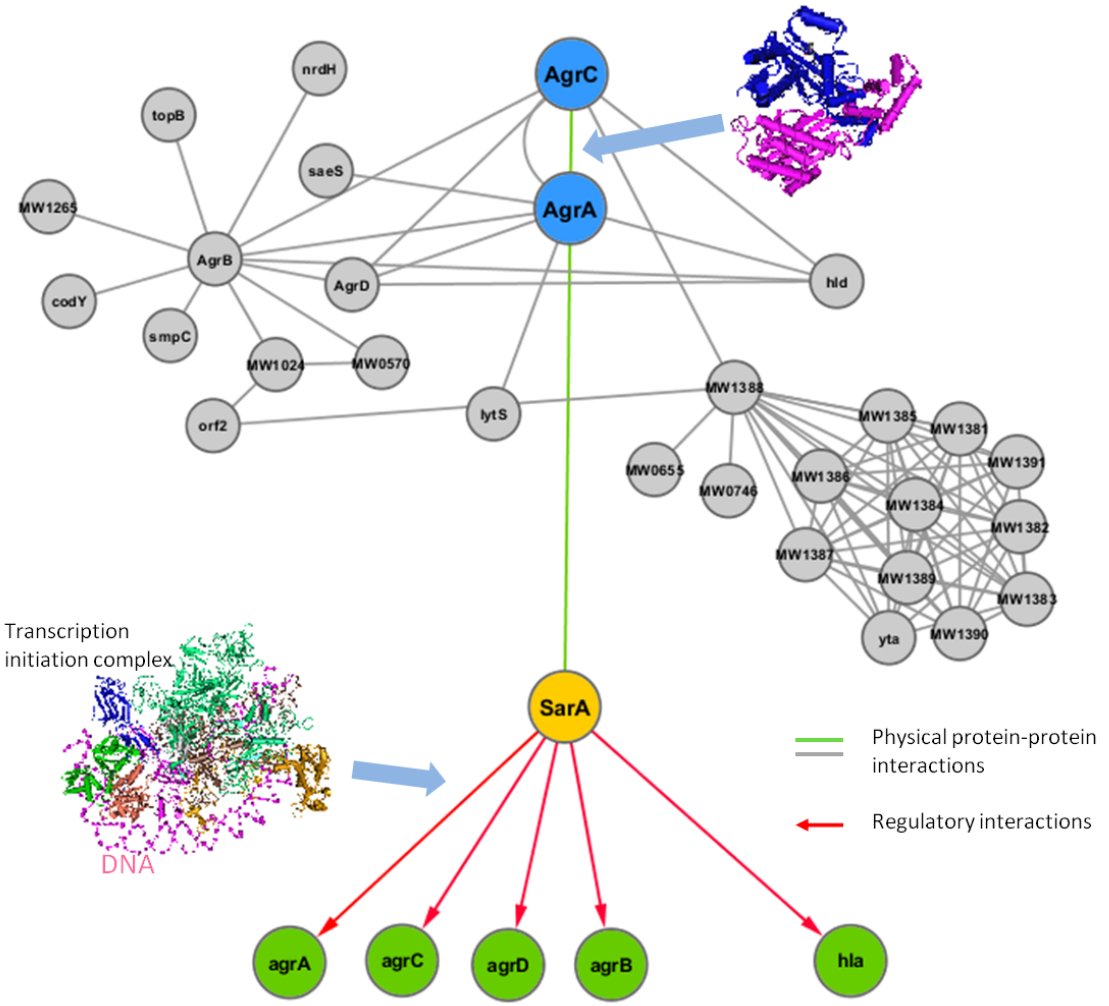

Figure 1. Example of a molecular network composed of regulatory protein-DNA and physical protein-protein interactions. The network displays a model of a two-component signal-transduction cascade in *Staphylococcus aureus*. The sensor histidine kinase receptor AgrC, upon binding of a cyclic peptid, phosphorylates the response regulator AgrA, which in turn activates the transcription factor SarA. In conjunction with AgrA, SarA induces the expression of Agr virulence factors and the hla gene (RNAIII). Nodes represent genes or proteins in this network and are colored depending of their role: target genes in green, transcription factor in orange, and the two component system in blue. Red edges represent protein-DNA interactions, whereas green or gray edges represent protein-protein interactions. The crystal structures representing the different type of interactions are taken from PDB and are only for illustration purpose (i.e. showing equivalent structures only).

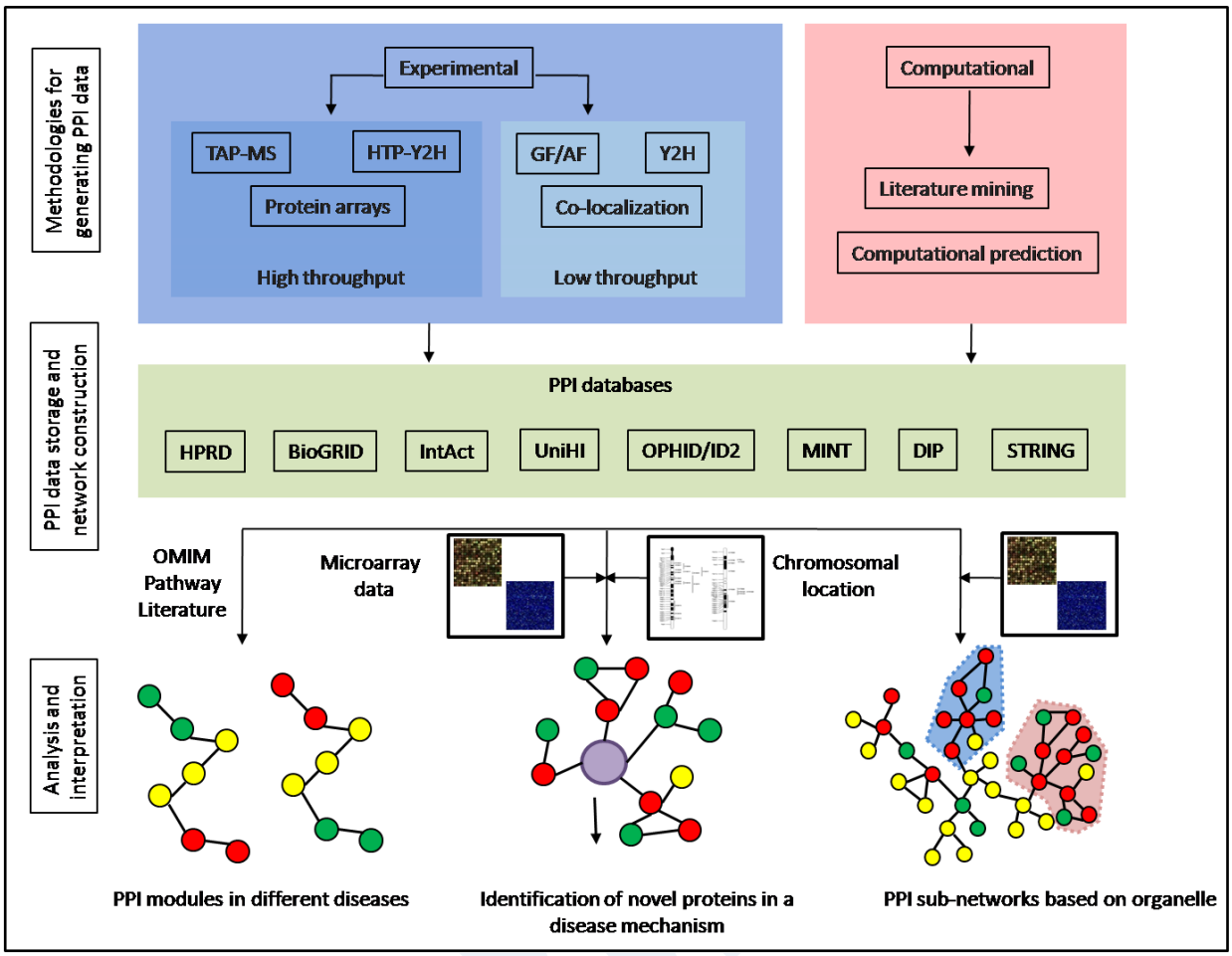

Figure 2: Overview of detection methods, databases and applications of PPI data.

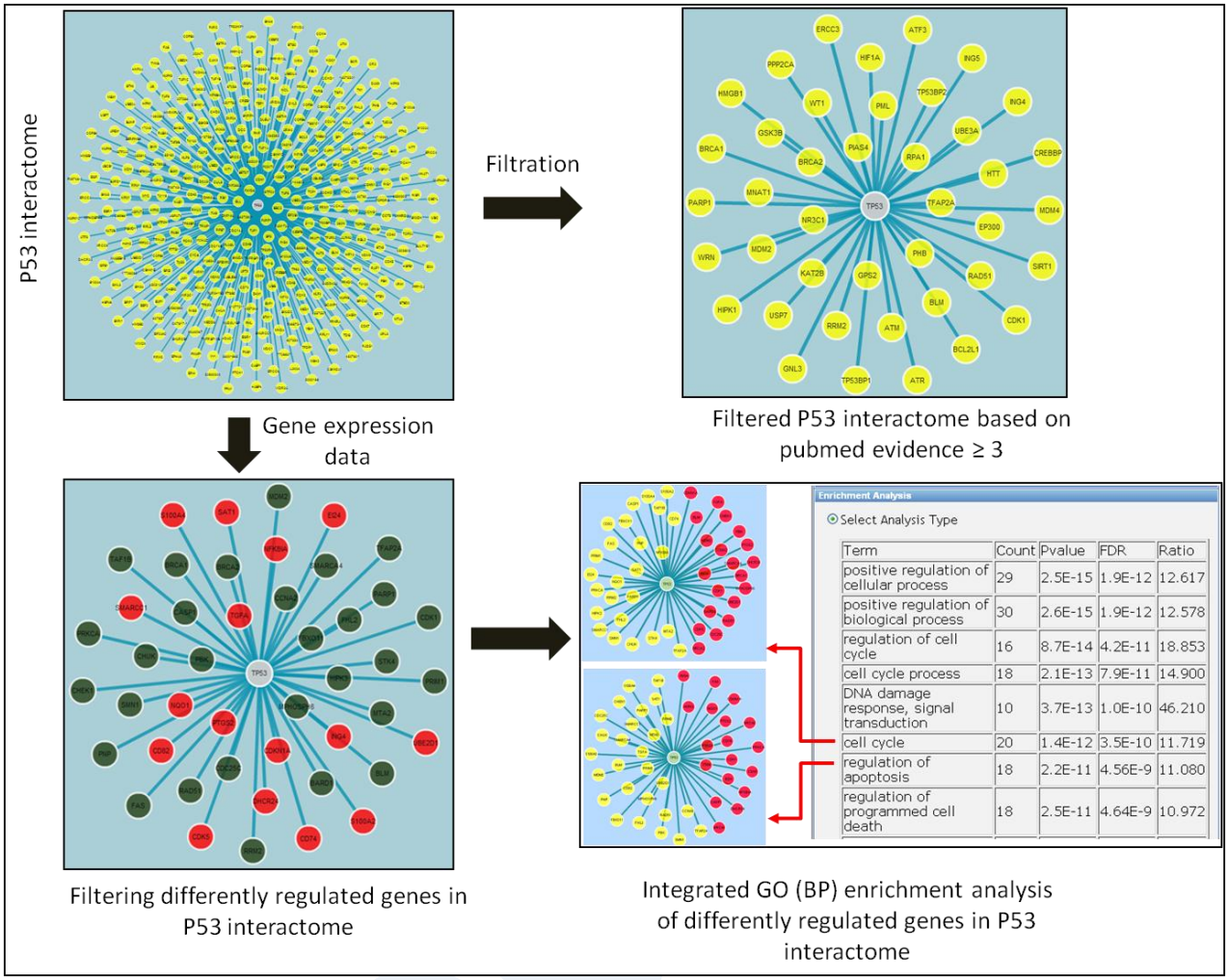

Figure 3: Example of network-based analysis using UniHI. For proteins of interest such as p53, interaction partners can be queried and visualized. The derived networks can subsequently be filtered based on evidence (e.g. number of PubMed references reporting the interaction) or based on gene expression data. All networks can be readily inspected for enrichment in biological processes using an integrated tool.