

## REVIEW

# Protein networking: insights into global functional organization of proteomes

Enrico Pieroni<sup>1</sup>, Sergio de la Fuente van Bentem<sup>2</sup>, Gianmaria Mancosu<sup>1</sup>, Enrico Capobianco<sup>1</sup>, Heribert Hirt<sup>2,3</sup> and Alberto de la Fuente<sup>1</sup>

<sup>1</sup> CRS4 Bioinformatica, c/o Parco Tecnologico POLARIS, Pula, Italy

<sup>2</sup> Department of Plant Molecular Biology, Max F. Perutz Laboratories, University of Vienna, Vienna, Austria

<sup>3</sup> Plant Genomics Research Unit, Unité de Recherche en Genomique Végétale (URGV), INRA/CNRS, Evry, France

The formulation of network models from global protein studies is essential to understand the functioning of organisms. Network models of the proteome enable the application of Complex Network Analysis, a quantitative framework to investigate large complex networks using techniques from graph theory, statistical physics, dynamical systems and other fields. This approach has provided many insights into the functional organization of the proteome so far and will likely continue to do so. Currently, several network concepts have emerged in the field of proteomics. It is important to highlight the differences between these concepts, since different representations allow different insights into functional organization. One such concept is the protein interaction network, which contains proteins as nodes and undirected edges representing the occurrence of binding in large-scale protein-protein interaction studies. A second concept is the protein-signaling network, in which the nodes correspond to levels of post-translationally modified forms of proteins and directed edges to causal effects through post-translational modification, such as phosphorylation. Several other network concepts were introduced for proteomics. Although all formulated as networks, the concepts represent widely different physical systems. Therefore caution should be taken when applying relevant topological analysis. We review recent literature formulating and analyzing such networks.

Received: August 7, 2007  
Revised: November 1, 2007  
Accepted: November 1, 2007

**Keywords:**

Complex networks / Interactomics / Network biology / Protein networks / Systems biology

## 1 Introduction

### 1.1 General remarks

Although large-scale high-throughput experimental techniques have greatly increased our knowledge, understanding the global organization of proteomes is still by far incom-

plete. A global view on the proteome is hampered by the complexity: there are tens of thousands of proteins and potentially hundreds of thousands of relations between them. Abstract representations of the proteome and the relationships are needed to be able to analyze and interpret such huge collections of data.

### 1.2 Why networks?

To understand living cells one must study them as systems rather than a collection of individual molecules. The study of systems consisting of thousands of interacting molecular species is very complicated and simplifying abstractions are necessary. The abstraction of intracellular processes into 'networks' is particularly fruitful [1, 2]. Networks provide a clear representation of complicated relationships between

---

**Correspondence:** Dr. Alberto de la Fuente, CRS4 Bioinformatica, c/o Parco Tecnologico POLARIS, Edificio 1, Loc. Piscina Manna 09010 Pula, Italy  
**E-mail:** alf@crs4.it  
**Fax:** +39-070-9243-4114

**Abbreviations:** PIN, protein-interaction network; PSN, protein-signaling network; SCC, strongly connected component; TAP, tandem affinity purification; Y2H, yeast two-hybrid

large numbers of elements and are used in scientific disciplines as diverse as sociology, epidemiology, molecular biology and physics. The network approach to complex systems has led to insights into evolution of networks and shed light on the interplay between structure and function. The main goal is to relate the structure, or ‘topology’, of networks to the biological function. Insights into the global topological organization of networks summarizing relationships between proteins will provide insights into functional organization of proteomes. Future advances will enable to understand complex diseases in terms of complex networks [3, 4] [see also dedicated sessions at the Pacific Symposium on Biocomputing (Pacific Symposium on Biocomputing, 2006, <http://psb.stanford.edu/psb-online/proceedings/psb06/#protein> and Pacific Symposium on Biocomputing, 2007, <http://psb.stanford.edu/psb-online/proceedings/psb07/#protein>)].

This review is meant to summarize and discuss the current status of network formulation and analysis in the field of proteomics. The goal of this review is to enlighten experimental proteomic researchers with concepts from Complex Network Analysis and to highlight the importance of formulating and analyzing networks. Therefore, we start out by introducing the basic concepts of Complex Network Analysis, a quantitative framework to investigate large complex networks using techniques from graph theory, statistical physics, dynamical systems and other fields. On the other hand, we would like to reach the community of Complex Network Analysts and make them appreciate the biological meaning of the networks in order to perform most effective analysis.

We discuss two main network models for proteomics. The first is the protein interaction network (PIN) (Fig. 1A), which summarize protein-protein binding events on a proteome-wide scale. PINs constitute the first network-oriented approach to proteomics resulting in a huge body of literature. The formulation of PINs opened doors to novel re-

search and insights into large-scale organization and evolution that can not simply be obtained without an explicit network perspective. We give an unambiguous definition for PINs. Experimental procedures to discover protein-protein binding interactions are reviewed and computational approaches for network fine-tuning using information from different data sources are discussed. To conclude this part we review the literature on Complex Network Analysis of PINs.

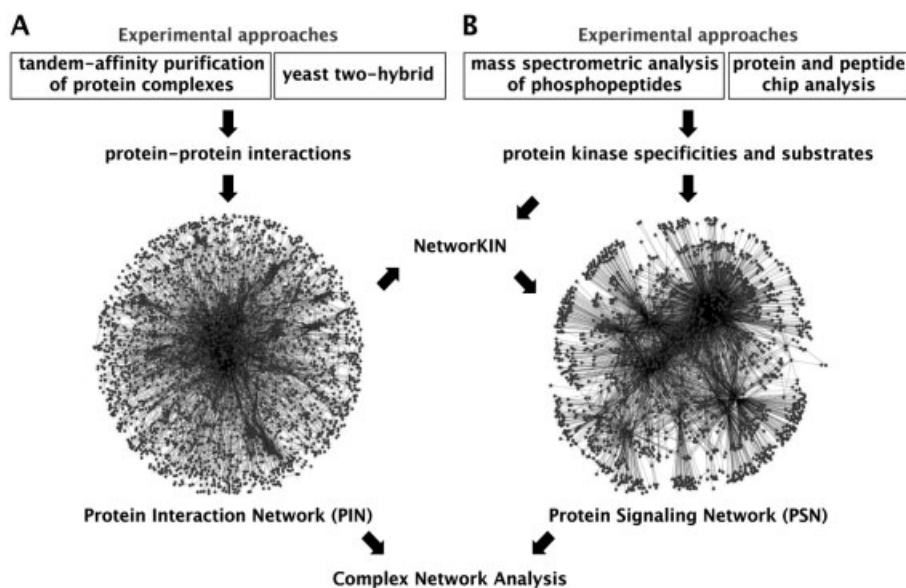
The second network model we define and discuss is the protein-signaling network (PSN) (Fig. 1B), in which the nodes correspond to levels of post-translationally modified forms of proteins and directed edges to causal effects through post-translational modification, such as phosphorylation. We review the current state of art in experimental techniques for high-throughput discovery of phosphorylation events and the formulation of PSNs. The application of tools from Complex Network Analysis to PSNs is not as extensive as for PINs, but this will change in the future as PSNs are more interesting than PINs in terms of information processing. We review results in this area and highlight biological insights resulting from a network-oriented perspective.

The review is concluded by describing other network concepts for proteomics that are expected to have a lower impact than PINs and PSNs for the understanding of the functional organization of living cells.

## 2 Introduction to complex networks analysis

### 2.1 Introduction to networks

Biological systems are complex, with many components (genes, proteins, proteins complexes, transcription factors, *etc.*) interacting and reciprocally regulating in an orchestrated



**Figure 1.** (A) Experimental approaches to formulate PINs. Nodes and undirected edges represent proteins and occurrence of binding between them, respectively. (B) Experimental approaches to formulate PSNs. Nodes and directed edges represent phosphoproteins and phosphorylation reactions, *i.e.* the effect (of a protein kinase) on the phosphorylation state of a protein, respectively. Both networks could be combined to enable Complex Network Analysis. Networks were drawn using Cytoscape [162]. The PIN is described in [76] and the PSN in [152].

way. At an abstract level we can simplify these systems and represent them as a collection of nodes, representing the interacting elements, connected by edges, representing the pair-wise interactions between the nodes. As effectively stated by Newman we have to answer the fundamental question “How can I tell what this network looks like, when I can’t actually look at it?” [5]. Complex Network Analysis precisely does this; it enables us to characterize the structure, or topology, of large complex networks. Below we give the basic terminology and concepts used in Complex Network Analysis. For a more in depth account we refer the interested reader to books and reviews: [5–11].

Nodes represent the system components, the variables, the actors. Nodes are graphically often depicted as small circles (Fig. 2). Edges represent certain relationships, or interactions, between the nodes, sometimes called ‘connections’, or ‘links’. Depending on the nature of the interaction, the edges may be directed (Fig. 2a), distinguishing between a source (or regulator) and a target (or regulated), or undirected (Fig. 2b). A network with directed edges is called a directed network, while one with undirected edges an undirected network. Directed edges are often depicted as arrows starting in the source node and ending in the target node. Undirected edges are simply lines drawn between two nodes. An edge can represent the presence of a relationship, but can also have an associated numerical value corresponding to the strength, or weight, of the relationship. A network carrying such numbers on the edges is called a weighted network.

It is also possible to associate a categorical variable to each link, called color, representing for instance the type of the interaction. The nodes can be of different kinds as well, for instance genes or proteins, and then can be themselves colored [5].

Networks can be represented graphically, but for analysis it is useful to describe them as matrices. The adjacency matrix is a square  $n \times n$  matrix, where  $n$  is the number of nodes, with entries  $(i, j)$  equal to 1 if there exists an edge from node  $i$  to node  $j$  and 0 otherwise. This matrix is typically very sparse for real world networks and is symmetric for undirected networks. For a weighted network the non-zero entries have real values instead of 1.

Two nodes connected by an edge are called adjacent or neighbors, the set of nodes adjacent to node  $i$  is called the neighborhood of  $i$ . A sequence of adjacent nodes is called a path. For directed network paths are directed and run along the edge directions. Directed networks are cyclic if there exists at least one directed path from a node back to itself, or acyclic if no such paths are present. Cliques are fully connected subsets of nodes where each node is adjacent to all others (Fig. 2).

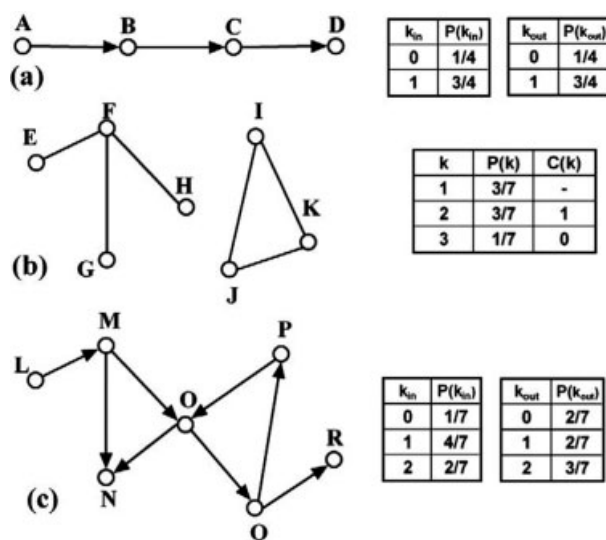
The weak component to which a node belongs is the set of nodes that can be reached from it by undirected paths. Large networks can have several separate components of which the largest component is usually subjected to analysis. For directed networks the division into components is more nuanced since the edge directions are taken into account. In the bow-tie representation [12–15] all nodes are assigned to

one of the following groups: strongly connected component (SCC) which contains nodes which can all reach each other through directed paths, in-component containing nodes that can reach the SCC through directed paths, but can not be reached from the SCC, and out-component containing nodes that can be reached by the SCC but can not reach the SCC through directed paths.

The first local characterization of a network is the node degree. The degree  $k$  of a node is simply the number of edges attached to it (Fig. 2). In the case of directed network we can distinguish between in-degree ( $k_{in}$ ), the count of incoming edges, and out-degree ( $k_{out}$ ), the count of outgoing edges. By averaging over all the nodes it is immediate to get the average degree.

By considering all node degrees we can obtain a global quantity, the degree distribution  $P(k)$ , which gives the percentages of nodes for each degree  $k$  (Fig. 2b). In the case of directed network, we can build the joint distribution  $P(k_{in}, k_{out})$  of having  $k_{in}$  in-edges and  $k_{out}$  out-edges (Figs. 2a and c). In the latter case, by summing the joint distribution on the in-degree we can recover the out-degree distribution itself, and vice versa. These distributions can be formally defined by summing on the columns or rows of the adjacency matrix [8]. The maximum degree is often a useful parameter, simply defined as the maximum of all node degrees.

Degree mixing is an important network feature, capturing how nodes with a particular degree interact with others nodes of particular degree. In assortative networks nodes with high degree tend to pair up with nodes with high degrees [16, 17], while networks in which nodes with high degree tend to pair up with nodes with low degrees are called disassortative. In the latter networks, the highly connected nodes seem to ‘repel’ each other.



**Figure 2.** Example of directed (a, c) and undirected networks (b). For each network the degree distribution is given, for undirected network (b) we also gave the clustering coefficient. Nodes I, J and K form a clique. The Figure is taken from [6].

Another important measure is the clustering coefficient, a measure of the network cohesiveness, that is how densely connected are the node and its neighborhood. In the case of undirected networks, the node clustering coefficient [18] of node  $i$  is defined as the number of edges between nodes adjacent to  $i$ , divided by the number of possible edges between them (Fig. 2). In other words, it quantifies how similar the neighborhood of node  $i$  is to a clique. The clustering coefficient  $C$ , a global measure, is then obtained by averaging the node clustering coefficients over all nodes. An alternative definition of clustering coefficient comes from social sciences, where it is defined as the ratio between the number of triangles in the network and the number of connected triples of nodes, divided by a factor of three to correct for over-counting of triangles. The difference between these two definitions is that the former tends to weight heavily the contribution of low-degree nodes [5]. In other words,  $C$  is the probability that two neighbors of a given node are themselves adjacent [5, 19].

Joining the two concepts of degree and clustering coefficient, we can define the clustering coefficient distribution,  $C(k)$ , as the average clustering coefficients of all nodes having degree  $k$ . In the case of directed networks the same definitions hold, simply using undirected version of the networks and the node degree  $k = k_{in} + k_{out}$ , however, it is also possible to distinguish between downstream and upstream contributions with respect to a specific node allowing to define the downstream and upstream clustering coefficient [20]. Again, the clustering coefficient can be formally defined by appropriate summing of the product of two adjacency matrix elements.

In the network, a naturally emerging concept is the shortest path length between two nodes, also called geodesic distance. It can be generalized to weighted networks as the minimum sum of weights along the path between two nodes. The largest shortest path length is defined as the network diameter. The diameter and the average geodesic distance then provide an estimate of the network overall *navigability*. To avoid problems with unconnected nodes the average path length can be defined as the harmonic (instead of arithmetic) average [5]. The shortest path length distribution,  $P(l)$ , is a third important global network feature, defined as the percentage of shortest path lengths of each size  $l$ .

Notice that while average degree, path length and clustering coefficient depend on the number of nodes and edges in the networks,  $P(k)$ ,  $C(k)$  and  $P(l)$  do not and could be used to capture generic features and thus classify and compare various networks [2].

Another important concept is centrality, which quantifies the topological importance of a node (or edge) in a network. Several centrality measures have been proposed [21]:

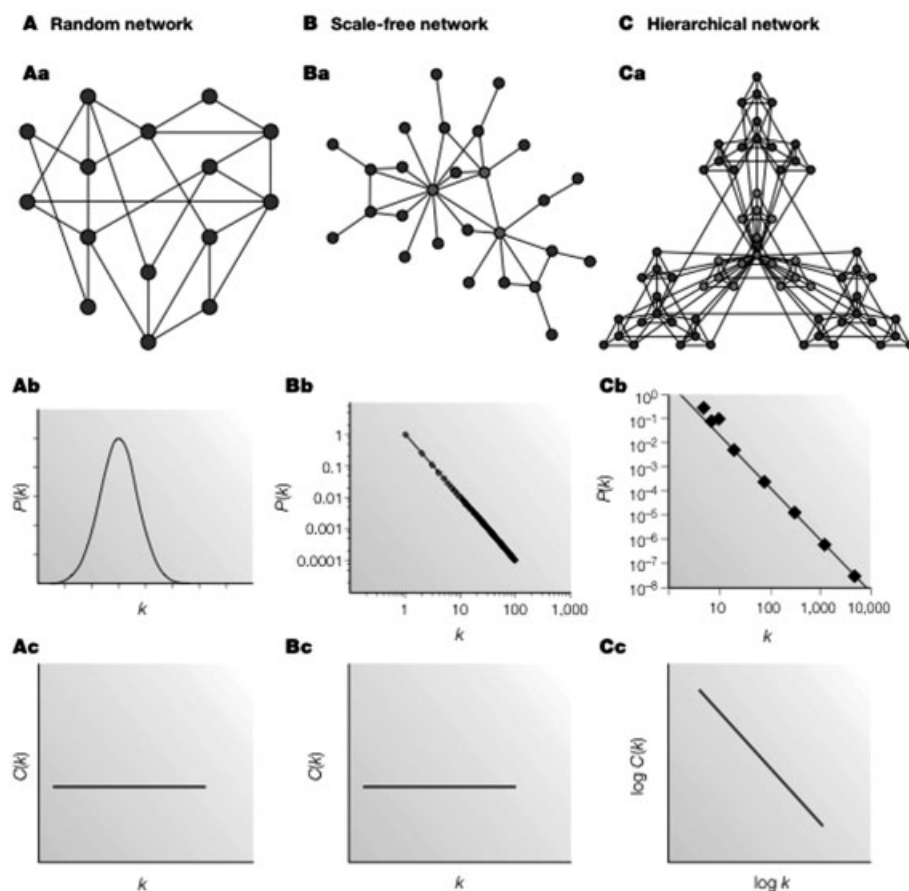
- (i) degree centrality: nodes with a large number of edges have high centrality;
- (ii) closeness centrality: nodes with short paths to all other nodes have high centrality;
- (iii) betweenness centrality: nodes (or edges) which occur in many of the shortest paths have high centrality.

## 2.2 Theoretical network models

Many theoretical network models have been proposed with the main goal of trying to capture features observed in real world networks. The first widely used model dates back to the pioneering work of Erdős and Renyi [22] and is called the random graph (Fig. 3A). In the remainder of this paper, we will refer to this model as the ER-network. To generate an ER-network given a fixed number of nodes,  $n$ , an edge between any two pair is iteratively added with probability  $p$ , until all possible distinct pairs have been taken into account ( $n(n-1)/2$  steps). Despite the simplicity of the model and the very few parameters ( $n, p$ ) this network is capable to show an impressive number of non-trivial behaviors and even mimic some features of real networks [23]. It is of course quite uniform or democratic: every node has the same average neighborhood. This statistical homogeneity is essentially the reason for which the degree distribution (that can be shown to follow a Poisson distribution) is peaked around the mean (Fig. 3Ab) and the clustering coefficient distribution is flat with a small average clustering coefficient (Fig. 3Ac). ER-networks are then quite well described by global average quantities. The shortest path distribution is peaked around small values and the average path is order of  $\log(n)$ , much smaller than  $n$ , an effect called “small world” [18]. Most real world networks seem to share such small world feature, likely due to the associated higher efficiency in the transfer of information or materials [18]. The first model capable to capture both the correct average shortest path and high, size-independent, clustering coefficient, was proposed by Watts and Strogatz [18]. This model is here referred to as the WS-network. The generating algorithm starts from a set of nodes regularly disposed on a lattice and then randomly rewires the edges with a fixed probability.

As for ER-networks, the degree distribution of the WS-network is also peaked around the mean value [24]. However, many degree distributions of real world networks have typically ‘fat tails’: they can have a few, but not negligible, number of nodes with degrees much higher than the average. In addition, many real world networks also show to be modular in structure, *i.e.* they contain certain distinguishable substructures. Even if ER-networks have a uniform character without hierarchical structures, some authors demonstrated they could nevertheless show high modularity, due to the fluctuations of the link formation process [25]. This fact is of paramount importance when assessing the statistical meaningfulness of modularity-based results on complex networks of whatever nature.

For many real world networks the degree distribution follows a power law  $P(k) \sim k^{-\alpha}$ , for some real positive  $\alpha$ , typically between 1 and 3. This behavior reflects the fact that most of the nodes have few edges, while only a few nodes, called hubs, have high degree [26]. These are the so-called scale-free networks (Fig. 3B), here denoted by Barabasi-Albert (BA-) networks, because there is no scale: the mean degree (scale) is not a good measure to characterize indi-



**Figure 3.** Example of random (A), scale-free (B) and hierarchical scale-free (C) undirected networks. For each network a pictorial representation (Aa, Ba, Ca), the degree distribution (Ab, Bb, Cb) and the clustering coefficient distribution (Ac, Bc, Cc) are given. The Figure is taken from [2].

vidual nodes (as it does in ER-networks). The dispersion (standard deviation) of  $P(k)$ , diverges for  $\alpha \leq 3$ , meaning that for  $\alpha > 3$  there are essentially no hubs, while for  $\alpha \leq 3$  hubs emerge, and the smaller the value of  $\alpha$ , the larger the hubs. In the case of  $\alpha < 3$  the average path length is order of  $\log(\log(n))$  much smaller than  $n$ , an ultra-small world property [2]. This kind of network can be grown using the principle of preferential attachment, in which the nodes are subsequently added to the network and are more likely to form links with higher degree nodes [26]. The starting configuration strongly influences the properties of the resulting networks [21]. Duplication and divergence models, in which individual nodes are occasionally copied and subsequently mutated with a certain probability, are more biologically motivated and can produce power law distributions as well [27]. Unfortunately, the average path length of the proposed models are too low compared to real networks and their clustering coefficient distribution is flat (Fig. 3Bc). In fact, many real world networks actually show a clustering coefficient distribution with power law tails,  $C(k) \sim k^{-\beta}$ , with  $\beta$  typically between 1 and 2, suggesting that lower degree node neighborhoods are highly cohesive; nodes with fewer edges tend to have higher clustering coefficients [28]. For many networks, a typical value of  $\beta = 1$  is the signature of hierarchical structure (Fig. 3C): sparsely connected nodes tend to

belong to highly clustered areas, which in turn are connected by a few internal hubs [2]. A simple model reproduces such properties: the starting point is a small cluster of  $p$  densely linked nodes, which is replicated a certain number  $q$  of times. Then, edges are added between the central nodes of each of the replicas and the original cluster. It has been shown that for suitable parameter choice ( $p = 4$ ,  $q = 3$ ) the model reproduced both power law exponent  $\alpha = 2.26$ ,  $\beta = 1$  and  $C = 0.6$  (size independent) [28]. Another important model is the geometric random network, generated by randomly placing nodes uniformly on a bounded grid, for instance a circle, and connecting two nodes only if their distance is less than a threshold.

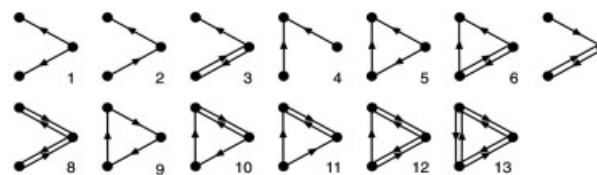
### 2.3 Sub-graph-based characterization

Motifs are small sub-graphs that are overrepresented in a network when compared to a null model [29]. The null model could be for instance a random graph [30] or, arguably better, a rewired version of the network under consideration with the same degree distribution. Motifs are sub-global topological features, linking the local organization to large-scale clustering properties, and are in no trivial way related to the clustering and degree distribution [31]. Motifs may provide insight into both the structure and function of regions of the

whole network, and even help to develop models for the evolution of biological networks [21]. Some authors believe that motifs may be seen as the atomic constituents of networks and thus can define universal classes of networks [32]. Motifs can be identified in directed as well as undirected networks. Obviously, there are many more directed subgraphs than undirected ones: for example, there are 13 unique directed 3-node motifs, while there are just two undirected (Fig. 4). Abundance of a given motif when compared to a reasonable null model is always an interesting signal, but one should be careful when relating such findings to functional biological aspects: which null model to use is still a controversial topic [30, 33, 34]. Furthermore, different types of networks may require different null models.

Local connection patterns can be used to classify and compare networks [32, 35]. For this purpose, Przulj *et al.* [35–37] proposed the concept of graphlet distribution as a powerful generalization of degree distribution. For instance, for a given node, they count the number of graphlets of the kind G1 (Fig. 5) the node is connected to. The node can link to G1 in two topologically distinguishable ways: to a central node or to a lateral node. Therefore, G1 represents two graphlets. In the same way, it is immediate to build all the 73 topologically distinct graphlets with two-to-five nodes, as shown in Fig. 5. Using this approach, two networks can be said to be similar if their graphlet distributions are alike.

Many authors [21, 38, 39] observed that most current research still focus on global network properties (average shortest path, clustering coefficient, assortativity, degree distribution, *etc.*), while most real networks are not homogeneous but have a clear modular structure. Modules can be determined in many different ways from the topology of networks [21]. One of the most recent and widely used techniques is based on modularity optimization [40, 41], in which the network is partitioned into modules in a way that maximizes the difference between the number of edges inside modules and the



**Figure 4.** Catalogue of all three-node motifs for directed networks. The Figure is taken from [29].

number of edges between modules. Quickly a debate emerged: first, it is surprising how many random graphs can present partitions with large modularity [8, 25]. Secondly, modularity optimization may fail to identify modules smaller than a typical scale that depends on the total number of links and on the degree of interconnectedness between modules [42].

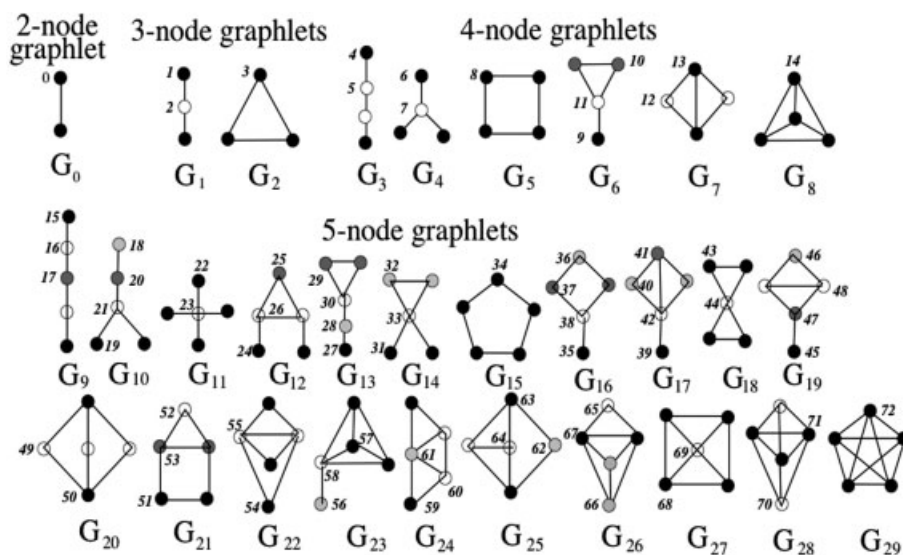
Other authors proposed a method that allows for screening multiple resolutions of the modular structure [43]. In this way, they abandoned the idea to maximize the modularity and thus find one static division in communities, but instead used the modularity as a detector to measure and access simultaneously to all the new scales of description of complex networks [43].

Complex Network Analysis provides a quantitative framework to understand different types of networks (Fig. 1). In order to be able to apply such tools in proteomics we need reliable representations of relationships between proteins as networks. One such representation is the PIN.

### 3 Protein interaction networks

#### 3.1 Introduction to protein interaction networks

We here define (consistently with many other authors) PINs as networks in which the nodes represent proteins and edges represent physical binding interactions between them. Two



**Figure 5.** Catalogue of all the 72 graphlets obtained for two-to-five node sub-graphs. The node arrangement is given by 29 pictures, called G0–G29. For each one of these connection patterns is possible to distinguish the node in the sub-graph to which a given node link. The resulting graphlets are enumerated from 0–71. The Figure is taken from [35].

proteins that were observed in an experiment to physically bind to each other will be connected by an undirected edge. It is important to note that several networks presented in current literature under the name PIN are actually not completely conforming to this definition (see Tandem affinity purification section below). PINs are sometimes referred to as ‘interactomes’ [44, 45] to indicate that they are collections of interactions at a proteome-wide scale. PINs have been compiled for a wide variety of organisms from all kingdoms of life, from bacteria such as *Escherichia coli* [46] to the yeast *Saccharomyces cerevisiae* [47, 48], from the fruit fly *Drosophila melanogaster* [49, 50] and the worm *Caenorhabditis elegans* [51] to the primate *Homo sapiens* [52–54]. The most predominantly used techniques for PIN formulation are yeast two-hybrid (Y2H) and tandem affinity purification-tagging (TAP) strategies. The first time a network was explicitly compiled from physical protein-protein interaction data was for yeast [55]. Since most experimental work and analysis results concern yeast we mostly focus on the yeast PINs. Several experimental and computational approaches to discover protein interactions have been described in the literature. We review these here and highlight the differences in the network representations.

### 3.2 Curated protein interaction databases

Several databases for protein interactions have been curated from the literature and are continuously updated. These include DIP [56–58], BIND [59–61], MIPS [62, 63], MINT [64, 65] and REACTOME [66, 67]. The overlap between the databases is very small [68, 69], making it difficult to obtain confidence in the interactions. On the other hand, it could be argued that each such database contains a different, slightly overlapping, sample of the entire network and that combining them would provide a better estimate of complete PINs. This idea may be supported by the fact that estimated sizes of PINs exceed the number of interactions currently stored in each of the databases [70, 71]. Most of the interactions in these databases are extracted from literature on ‘small-scale experiments’ (as opposed to ‘high-throughput experiments’). While in general discoveries in small-scale experiments are assumed to be of better quality than those by high-throughput experiments one could argue that the opposite is true: high-throughput experiments require extensive standardization and calibration, while each small-scale experiment is performed differently each time. Furthermore, in contrast to small-scale experiments, in which most of the focus is on subsets of the proteome, *i.e.* the proteins considered ‘interesting’ by researchers, the high-throughput experiments give an unbiased view on the proteome. This then leads to a higher confidence in the PINs obtained by high-throughput means rather than those obtained from the currently available curated databases.

### 3.3 The yeast two-hybrid system

The yeast two-hybrid system (Y2H) is a method to test pairwise protein-protein interactions [72] and has been used for nearly two decades [73]. The system has been employed for high-throughput discovery of protein interactions [47, 48, 51]. The technique allows the detection of an interaction between a bait protein, which is fused to the DNA binding domain of the Gal4 transcription factor, and a prey protein that is fused to the transcription activation domain of Gal4. An interaction between the bait and prey proteins reconstitute proximity of the separate Gal4 domains and restore Gal4 function. The output of the interaction is the Gal4-dependent activation of several reporter genes, and nuclear localization signals are included in the fusion proteins to allow the interaction to take place in the nucleus. Of course, forcing two proteins together will give rise to a high false-positive rate, in the sense that although these proteins truly physically bind they will never do so inside cells, because of different localization, or because they are never simultaneously expressed. False negatives may occur because PTMs crucial for interaction might be lacking (for instance between phosphoproteins and phosphoprotein-binding domains) by localizing the hybrid proteins in the nucleus and by expressing non-yeast proteins in yeast. Most results from Complex Network Analysis (see below) are obtained from two yeast PIN obtained by Y2H.

### 3.4 Tandem affinity purification of protein complexes

TAP is a more recently established technique to purify protein complexes. The TAP technology has allowed the dissection of hundreds of protein complexes from yeast [74–76]. In contrast to the Y2H system, the TAP method enables the elucidation of native protein complexes (if not disturbed by the TAP tag itself) by pulling down a TAP-tagged bait protein from cell extracts and determining its co-purifying partners by MS. Although no comprehensive TAP purification strategy towards animal or plant PINs has been undertaken, improvements of the TAP tag for purification of TAP complexes from these organisms [77–79] and the development of highly sensitive and accurate mass spectrometers will allow such analysis in the near future.

The networks obtained from TAP studies are different from the PINs as defined above. This is because the authors assume edges between the bait and any other protein that is co-purified with it. This way, proteins within the same complex will be joined by edges, while this does not necessarily mean direct physical binding between them. For example, if bait A co-purifies B and C, but A only directly binds B which in turn binds C there will be an interaction between A and C which does not correspond to a direct physical binding. It was shown that computational discovery of protein complexes from TAP-derived networks is more accurate than from Y2H-derived networks [76] by comparing predicted complexes to the ones present in the MIPS database. This is

expected because the TAP-derived networks explicitly include information about protein complexes through the additional indirect edges. While for this purpose TAP-derived networks are superior, investigations into the large-scale organization of the proteome requires networks that reflect precisely the ‘wiring’ structure of physical binding, *i.e.* PINs such as defined above, with only edges that correspond to direct physical binding. Collins *et al.* [80] combined two important TAP datasets to obtain a high confidence network of 1622 nodes and 9074 edges. Pu *et al.* [81] showed that protein-complex detection from this network occurred with highest reliability as compared to other datasets.

### 3.5 Protein and peptide chips for proteomic research

Powerful alternatives to Y2H and TAP methods for studying PINs are peptide and protein chips. They consist of arrays of up to thousands of peptide or proteins individually spotted onto a carrier such as a glass slide. Protein and peptide chip experiments allow the quantitative assessment of PINs by applying prey proteins or peptides on the chip and measuring the binding affinities to each of the bait proteins or peptides on the chip [82]. Protein and peptide arrays can also be used for many other purposes (for instance discovery of protein kinase substrates, see below). The major drawback is the lack of physiological context in this *in vitro* approach.

### 3.6 Probabilistic models and data integration

Protein interaction data present a variable degree of reliability. PINs are expected to be largely incomplete and to contain a number of incorrect edges [44, 69, 83–85]. For each detected interaction, investigating its inherent reliability relies on the definition of a gold standard [86], *i.e.* a reference set of true-positive interactions – a set of interactions that is assumed to be ‘real’ – and a set of true-negative interactions – a set of interactions that is assumed to be absent. The gold standard dataset can then be used to optimize the performance of computational methods for reliable prediction of PINs. To improve coverage and accuracy it is necessary to combine and incorporate heterogeneous sources of information. Such information includes gene expression data [87–89], knockout phenotypes, subcellular localization, genetic interactions and phylogenetic profiles [90] and Gene Ontology. STRING (SearchTool for the Retrieval of Interacting Genes/Proteins) [91–93] is a database that offers a mix of known, predicted and transferred interactions covering many organisms, also those not (yet) experimentally addressed by high throughput analysis. The reliability of the interactions is also determined by the assignment of a confidence score, where the information sources are gene co-expression, automated text mining and genomic location. A score delivers the confidence gained from association (in Naive Bayes style) of various evidence sources, which are ‘naively’ considered independent on each other, and is calculated as a combined expression of scores for individual instances  $S_i$  of

evidence:  $S = 1 - \prod_i (1 - S_i)$ . While each type of evidence alone is not sufficient, the integration of several sources of evidence strongly improves predictions of interactions [88].

## 4 Complex Networks Analysis of PINs

In the introductory part, we have illustrated several characteristics that can be observed in networks by applying techniques from Complex Network Analysis. It is worth to point out that we have defined such methods and properties for both undirected and directed networks. Below, we describe findings on the application of the analysis to PINs, which are intrinsically undirected networks, as edges are binding relationships between proteins: there is neither flow of information nor mass between nodes – an edge simply indicates that two proteins bind. As a consequence, one should be careful when applying measures based on distance in the network, such as ‘path lengths’ and related properties, since these could be completely abstract, not allowing for any physical interpretation. The underlying assumption when considering measures involving distance is that an undirected edge between protein A and B corresponds to two directed edges, one from A to B and one from B to A. This assumption of bi-directional flows is incorrect for PINs. Even if some binding events may be accompanied by signal flows (see Section 5), this is not true in general. As shown below, there are several works that investigated PINs using network measures involving paths. However, discovered relationships between such measures and biological properties could be simply due to other network measures that are truly related to those biological properties and correlate with the distance-based measures. Many networks measures are related – if one measure is ‘high’ in a network a related measure is always high as well – but often have different physical interpretation. Knowing the physical nature of PINs should help in selecting the relevant network measures.

### 4.1 Degree distributions

Several authors have shown that the degree distributions of most PINs are well fit by a power law, indicating that these are scale-free networks in which most proteins have a small number of neighbors while a small number of proteins are hubs; they have a large number of neighbors [69, 84, 94, 95]. Others have found a slightly faster decaying tail, *i.e.* a power law with exponential decay [80, 96] that shows fewer and smaller hubs than a pure power law would do. If PINs are scale free is thus not clear. In addition, there is currently a hot discussion about the interpretation of the power law observed in the degree distribution of most of real world data. The point in discussion is that real world data are noisy and inaccurate (particularly for the higher degree), incomplete and data are ‘sampled’ from a potentially much wider network. To assess the validity of the power law findings, some authors

demonstrated that sampling from a scale-free network could result in a non-scale-free network [97]. More importantly, it was shown that a power law tail could be observed in networks obtained by sampling from networks having degree-distributions very distinct from power laws! [84]. More precisely, these authors generated four theoretical interaction networks with quite different topologies (random, exponential, power law, truncated normal). A partial sampling of these networks resulted in sub-networks with topological characteristics that were virtually indistinguishable from those of current (partial) PINs. Their conclusion was that, with the current limited coverage levels, the observed scale-free topology of existing PINs could not be confidently extrapolated to complete PINs. Still, they pointed out that it is more likely that the current results are due to the fact that complete PINs are truly scale-free rather than having other degree distributions (see also [69]). The scale-free distribution is not as sensitive to false positives (erroneous links) in the network as they are to false negative (missing links) [98].

Purely scale-free or not, fact is that there are hubs with many more edges than the average degree. It has been computationally shown that networks with scale free degree distributions are more robust towards random node removal than ER-networks, and more sensitive to targeted attacks of the high-degree nodes [99]. This observation provides a link between network topology and the phenomenon of robustness of biological systems. This then suggests that highly connected nodes in PINs are more important than lowly connected nodes. Indeed, Jeong *et al.* [96] showed a positive correlation, though not very large, between node degree and lethality in yeast PIN obtained mostly by Y2H experiments. Knockout mutants missing a gene coding for a high-degree protein were lethal with higher probability than low-degree protein knockout mutants, indicating that hubs indeed play an important physiological role. Other authors [100] showed the existence of a small positive correlation between betweenness centrality and lethality in yeast PIN obtained by combining interactions from the curated databases DIP and MIPS. However, since this measure involves the concept of paths, and paths do not physically exist in PINs, the relationship must be due to a confounding correlation with another network property, such as the degree centrality. However, interestingly they found a certain number of proteins with high betweenness centrality, but low degree. Indeed, no significant correlation between that class and lethality could be found [100], indicating that it is the degree that matters. Han *et al.* [101] proposed that there are two types of hubs: (i) 'party hubs', whose genes are co-expressed with all their neighbors' genes over many physiological conditions, and (ii) 'date hubs' whose genes are co-expressed with only one or few neighbors' genes in each physiological condition. The latter are thus not true hubs since their degree is low and depends on the physiological state. Other authors have disputed the existence of these categories of hubs [102, 103].

## 4.2 Node degree correlations

In a pioneer work, Maslov and Sneppen [104, 105] considered yeast PIN from Y2H data consistent of 4549 edges between 3278 proteins. They quantified the correlations between degrees of the nodes and compared these to a null model, in which all links were randomly rewired. They observed what is called 'disassortative mixing': links between highly connected proteins were systematically suppressed, whereas those between a highly-connected and low-connected pairs were highly favored. As originally stated by the authors, this effect could have a clear biological meaning: confusing cross talk between different functional modules is much less likely. Moreover, hubs tend not to share their neighbors with other hubs. This may increase the overall robustness by localizing the effects of deleterious perturbations around the hub where disturbs generated [105]. This anti-correlation then provides a certain degree of protection against such attacks. This may also explain why the correlation between the degree of a given protein and the lethality of the mutant cell lacking this protein is not particularly strong [105]. An alternative explanation of these findings is that the hubs act as important central compounds of complexes by holding many proteins together, do not bind to other complex-centers. In addition, few proteins are shared between complexes.

## 4.3 Hierarchical topology

The hierarchical structure of PIN is extensively investigated (see for instance [69]). Mainly, the idea is that proteins with similar function should be segregated in clusters clearly separated by other proteins. These authors analyzed four different PIN: two based on Y2H datasets and two on curated databases (MIPS and DIP). All networks showed a hierarchical structure, sustained by scale free topology with a hierarchical modularity as evidenced by decaying cluster coefficient for increasing degree. The clustering coefficient was compared with a properly defined segregation parameter, finding that for some functional classes (*e.g.* cellular communication) proteins stay close together, with a small clustering: proteins interacts with each other but not in a strict way. Other classes, instead, (*e.g.* cellular organization) tend to stay clustered together. A further characteristic of neighbor proteins in PINs is that they tend to be localized in the same cell region, so that the topology of the network reflects the cell's physical compartmentalization (*e.g.* edges between proteins belonging to the mitochondrial matrix are 100 times more probable than by chance). It may then be possible to predict the function of a protein based on its position in the network [106]. For example, if a protein with unknown function has many neighbors with a particular Gene Ontology classification, it is likely to belong to that class as well [55, 107].

#### 4.4 Sub-graph analysis

An exciting result coming from the study of specific sub-graphs in yeast PIN is the work of Wuchty *et al.* [108]. They showed that specific sub-graphs contain more conserved proteins than by chance. They identified highly conserved proteins by using InParanoid, a database of orthologs [109], and considering conserved all the yeast proteins with an ortholog in all five eukaryotes (*H. sapiens*, *A. thaliana*, *C. elegans*, *M. musculus*, *D. melanogaster*). This can be considered as suggestive evidence of the functional biological role of these small sub-graphs, because evolution preserves modules with specific biological function [110].

Another application of small sub-graphs is proposed by Przulj *et al.* [35, 36] where the authors defined a similarity measure of two networks, essentially based on the 73 graphlet distribution functions, collapsing all these degree of freedom into a single number. In this way they were able to show that almost all of the considered 14 eukaryotic PINs are better modeled by 3-D geometric random graphs than by either ER-networks, BA-networks or hierarchical networks.

#### 4.5 Modular structure and multi-protein complexes

PINs have been shown to have modular structure [111]. An immediate physical interpretation of such modules is that they are multi-protein complexes, but there may be modular structure reflecting additional functional properties of the network [38].

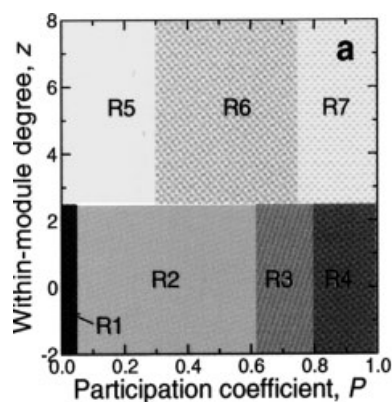
One way to detect protein complexes in PINs are so-called 'clustering algorithms' or 'community detection algorithms'. Several of such methods have been developed and recently have been evaluated [112]. Four algorithms, MCL: Markov Clustering [113, 114], RNSC: Restricted neighborhood search clustering [115], SPC: super paramagnetic clustering [116] and MCODE: molecular complex detection [117], were compared in their ability to 'rediscover' the annotated protein complexes listed in the MIPS database. The MCL algorithm simulates a flow on the network, and at each iteration an inflation step is applied to enhance the contrast between strong/weak flow regions. This process converges toward a partition in modules. On the other hand, RNSC is a local optimization search algorithm simply exploring the solution space and attempting to minimize a specific cost function. This objective function is based on the number of intra-clusters and inter-cluster edges. Starting from an initial random state the algorithm subsequently moves a node from a cluster to another, retaining the move if it reduces the cost. To evaluate the robustness of the algorithms to both false positives and negatives, these authors constructed modified networks by randomly removing or adding edges to the original network. Each clustering algorithm was then applied to the altered networks with various parameter settings, and the resulting clusters were compared with annotated complexes. They concluded that MCL is remarkably robust to alteration, while RNSC is more sensitive to edge deletion but less sen-

sitive to the use of suboptimal parameter values. The other two algorithms (SPC and MCODE) performed remarkably weaker over all aspects. A recent algorithm has been developed for pattern recognition and proposed but not yet extensively benchmarked for biological networks [118, 119]. The authors defined a similarity measure between pairs of nodes, then, starting with an initial random sub-network, the nodes exchange information about their similarity with their neighbors. The process proceeds and gradually a set of nodes with higher affinity emerges. The authors claim that the result is quick, accurate and less sensitive to a good choice of the initial data subset. Even if not yet formally compared to MCL, this process strikingly resembles the network flow of the MCL algorithm.

Other authors used the modular structure of networks [38] to assign a 'role' to nodes according to two main measures:

- (i)  $z$ , the relative within-module degree, measuring how well a node is connected to the other nodes inside the module
- (ii)  $P$ , the participation coefficient, measuring how well the node is connected to different modules.

Looking the distribution of ( $z, P$ ) values in real networks, the authors envisaged an interesting and useful classification of nodes (Fig. 6). The first subdivision is between hubs (R5, R6, R7 regions of the  $zP$  plane), having high values of within-module connectivity  $z$ , and non-hubs (R1–R4 regions), having small values of  $z$ . Then, the nodes in both these categories are further subdivided with respect to their capability to bridge to other modules: (R1) ultraperipheral-nodes, characterized by very small values of the participation coefficient; (R2) peripheral nodes, still with small values of  $P$ . Subsequently, participation increases and nodes start to have more connections: satellite connector nodes (R3), and finally nodes, still not being hubs, are strongly linked to other



**Figure 6.** Definition of seven regions on the ( $z, P$ ) plane, enumerated R1–R7. R1–R4 corresponds to non-hubs, which are weakly connected inside the module, while R5–R7 corresponds to intra-modulus highly connected hubs. Depending on the value of  $P$ , a further subdivision is done in ultra-peripheral (R1), peripheral (R2), satellite (R3), kinless (R4) nodes and provincial (R5), connector (R6) and global (R7) hubs. The Figure is taken from [38].

modules, kinless nodes (R4). The same reasoning applies to hubs, distinguishing between provincial hubs (R5), connector hubs (R6) and global hubs (R7). Their results show that PINs of yeast and worm are disassortative (as shown before [105] for the yeast), but the repulsion is limited to between hubs of type R6 and between R5 and R6, and it does not affect global hubs (R7) at all [38]!

The topologies of PINs have been studied extensively and we presented some of the important results. We now switch to PSNs, another complex network model for proteomics for which recently examples emerged in the literature.

## 5 Protein signaling networks

### 5.1 Introduction to protein-signaling networks

Although large-scale high-throughput experimental techniques have greatly increased our knowledge, our understanding of signal processing by cells is still by far incomplete. Multiple PTMs can transform each protein in the proteome into a dynamic and multifunctional unit [120]. Most studies on signaling networks have focused on one particular PTM to decrease complexity. Evidently, combining of data sets from different large-scale approaches will enhance the construction of entire signaling networks. Molecular networks have been constructed based on physical and functional interactions [121–123]. Large-scale analysis revealed signaling events that underlie apoptosis on a systems level [124]. Signal transduction pathways can be modeled at different levels of detail [125, 126] ranging from detailed mathematical models to graphical representations. From such networks, novel therapeutic strategies could be envisaged [127].

Several mathematical models based on ordinary differential equations have been formulated and their parameters optimized to fit experimental observations [128–131]. While studies with such models provide many insights into the dynamics and function of signal transduction pathways, formulating such detailed models is a difficult problem requiring a huge amount of experimental data, which is not commonly available, certainly not at a proteome-wide scale. The first requirement of such a modeling approach is the knowledge of the pathway structure, *i.e.* which are the targets of kinases, phosphatases, *etc.* and which reactions are involved. Inferring interaction structure at the proteome wide scale requires an abstraction of signal transduction pathways into PSNs.

We define (consistently with other authors) PSNs as networks in which the nodes correspond to levels of post-translationally modified states of proteins and directed edges to causal effects, indicating that the post-translationally modified state of one protein changes the post-translationally modified state of another. Nodes thus represent quantitative variables, *i.e.* concentrations of the post-translationally modified states. A wide variety of PTMs have been discovered, of which phosphorylation is the most studied one [132]. Source

nodes in PSNs will often be kinases with activating edges pointing out of them, but note that phosphatases (which reduce the level of the phosphorylated state of proteins) could be presented by inhibiting nodes. In PSNs no reactions appear like in the classical diagrams depicting signal transduction pathways. The networks described below almost exclusively involve protein phosphorylation. Ultimately, all PTMs will be included in PSNs as complete models for functional regulation of proteomes.

### 5.2 Perturbation strategies

Two recent studies outline how PSNs can be obtained *in vivo* through quantitative experimentation and perturbation analysis. The general idea behind those approaches is simple: components of the system are perturbed (in concentration or activity) and responses of the other components are measured. In this way causal-effect relations can be established, but in a next step one has to distinguish between direct and indirect effects [133]. In a PSN the edges only represent direct causal effects. Santos *et al.* [134] show a proof of principle on a small network of three interacting human mitogen-activated protein (MAP) kinases (MAPKKK, MAPKK and MAPK). These authors employed a perturbation strategy initially proposed to infer the structure of Gene Networks [135, 136] and later adapted for signaling networks [137, 138]. Perturbing the concentration of each of the kinases by RNA interference (RNAi) and measuring the response of the other kinases enabled to solve the interaction structure using a linear algebra approach [135–138]. Interestingly, they could show that the network structure differed upon stimulation by different hormones.

A statistically sound approach is outlined in Sachs *et al.* [139] who studied a signaling network of 11 proteins. In their approach the systems' components are specifically perturbed and responses are measured in a large number of replicates (each replicate about 700–900 times) on a single-cell level [139]. Then Bayesian networks are employed to identify the best network model fitting all perturbation data. Comparing the inferred network to the known pathway it was concluded that the inference was highly reliable. The approach was unable to detect the feedback loops owing to the inability of Bayesian networks to discover cyclic dependencies.

### 5.3 Phosphoproteomics

Novel methods for phosphopeptide isolation combined with mass spectrometric identification of phosphopeptide sequences now enable thousands of phosphorylation sites to be mapped [140, 141]. Quantitative MS-based methods have enabled the measurement of changes of individual phosphorylation sites during a time-course of a particular treatment or during different treatments [140, 142–144]. The resolution is at the site of phosphorylation and each phosphoprotein has on average at least three sites that are phosphorylated [140]. An alternative network representation

could thus involve individual sites as nodes rather than proteins to allow the connection of individual kinases to specific sites.

The above strategies enable to discover the *in vivo* active PSNs. Only relationships that are dynamically active in the used experimental condition can be discovered and, as evidenced in Santos *et al.* [134], the structure of PSNs can widely vary between physiological conditions. Below, we describe *in vitro* and *in silico* techniques that allow for identification of all potential interactions in PSNs.

#### 5.4 Chips for protein phosphorylation measurements

As mentioned above, peptide and protein arrays can be used as an approach to obtain PSNs. Both have been used to determine the substrate specificity of recombinant yeast protein kinases [145, 146]. Using yeast proteome chips, Ptacek *et al.* [145] found that highly related protein kinases phosphorylated different sets of proteins, suggesting that chips are useful tools to identify specific protein kinase substrates. By testing 87 of the 122 potential yeast protein kinases, 1325 of the 4400 proteins on the array were phosphorylated. Because recombinant kinases are often inactive in absence of their natural activators, kinases can also be applied on the chip as active complexes of different proteins [145]. However, as contextual information is lacking (see below), protein chips are only predictors of potential kinase-substrate connections. In addition, the absence of essential scaffolds or activating signals inevitably leads to false negatives in this approach. False positives might be caused by bringing a kinase artificially close to a substrate that it will never meet in its natural environment. Moreover, protein chips do not provide site-specific information.

Peptide chips have been used to determine proteome-wide kinase activities in animal and plant cell extracts or purified kinases, measured by the incorporation of radioactive ATP or by using phospho-specific antibodies [147–150]. This technique enables quantitative, high-throughput analysis of kinase activities in extracts of cells subjected to a range of conditions against a large number of known *in vivo* phosphosites [147]. In addition to the false negatives and positives envisaged for protein chips, peptide chips have the additional drawback that essential docking domains spatially separated from the phosphosite may be lacking. This is crucial since kinases such as MAPK specifically bind to their substrates *via* docking domains that can be located more than 100 amino acids away from the phosphorylation site.

#### 5.5 Computational discovery of PSNs

High-throughput, peptide-based methods allow the screening for phosphomotifs (conserved sequences of amino acids around phosphosites) of individual protein kinases [151]. However, because often multiple kinases share specificities towards peptides *in vitro*, knowing the phosphomotifs is not enough to couple kinases to phosphorylation sites in sub-

strates. By a novel approach termed NetworKIN, Linding *et al.* [152] discovered novel PSN by combining knowledge on phosphomotifs with contextual information provided by the STRING network [91–93] (see Section 3.6). Such network information determines at least 60% of kinase specificities, demonstrating its importance for modeling cellular systems [152]. Using only kinase consensus motifs gives a low prediction accuracy, but incorporation of contextual (network) information increases the accuracy by 2.5-fold [152]. The resulting prediction accuracy of more than 60% provides a solid ground for analysis of individual kinase-substrate pairs and for investigations of the global topology underlying signal processing in human cells (de la Fuente, A., Fotia, G., Maggio, F., Mancosu, G., Pieroni, E., Insights into biological information processing: structural and dynamical analysis of a Human Protein Signalling Network. Submitted to *Journal of Physics A* 2008). Indeed, Linding *et al.* [152] could verify several predictions by showing novel edges between kinases and substrates within the DNA damage pathway. With further improvement of the STRING resource, accuracy will certainly further increase. In addition to direct protein-protein interactions, STRING also provides indirect protein or genetic interactions, which is important since scaffold proteins play active roles to fine-tune the output of signaling cascades [153]. Including information on PINs to construct PSNs is expected to reveal connections that otherwise would not be found. Indeed, PINs and PSNs largely overlap in the case of kinases and their substrates [145, 154].

Combining peptide and protein chip experiments and the NetworKIN algorithm to connect protein kinases with their potential substrates and quantitative MS-based methods to enable site-specific phosphorylation profiling in time might allow, ultimately, the construction of dynamic PSNs.

## 6 Complex Networks Analysis of PSNs

Most tools from Complex Network Analysis have been developed for undirected networks. Sometimes directed networks are analyzed ignoring the directions, as if they were undirected. While this simply enables the application of the tools for undirected networks, one has to be very careful, since often it is silently assumed that an undirected edge establishes communication in both directions, thus representing a directed edge in both directions. This is of course a wrong assumption, as in PSNs a clear direction of signal flow is defined. Furthermore, ignoring the knowledge of direction is a loss of information. Most concepts for undirected networks can be straightforwardly adapted for directed networks. For example, as mentioned above, the concept of degree distributions can be extended to in- and out-degree distributions and the concept of cluster coefficients can be extended to up- and down-stream clustering [20]. Such distinctions are crucial since ‘hubs’ with only outgoing edges will be functionally completely different from hubs with only incoming edges or nodes with a high number of both. PSNs

have not been subjected to Complex Network Analysis at the same extent as PINs. The work of Linding *et al.* [152] resulted in a human PSN of 1810 nodes and 5189 edges. For this PSN it was shown that the degree distribution again followed the familiar power law. Furthermore it was shown that the clustering coefficient of nodes decreased with their connectivity indicating a hierarchical structure. In their topological analysis they did not take edge directions into account in order to be able to compare the networks properties of the PSN with those of the undirected networks used in the construction process, *i.e.* the PIN and context network. Most insights into PSNs will of course be obtained when taking directions into account. We expect that soon a large body of literature will appear on Complex Network Analysis of directed PSNs (de la Fuente, A., Fotia, G., Maggio, F., Mancosu, G., Pieroni, E., Insights into biological information processing: structural and dynamical analysis of a Human Protein Signalling Network. Submitted to *Journal of Physics A* 2008).

## 7 Additional protein network models

A Protein Homology Network (PHN) is a network in which nodes are proteins from potentially multiple organisms and edges between them are drawn based on a certain degree of sequence homology. Starting from 251 prokaryotic genomes, a PHN of 633 404 nodes was compiled [155] by performing DNA sequence similarity comparison and linking each pair of proteins that exceeded a given similarity threshold. 127 856 proteins resulted isolated, while the others were classified using a modularity optimization into 28 226 PHN-families containing at least two proteins. The largest component identified has 39 321 nodes and  $4.4 \times 10^6$  links and showed a clear modular structure. Using such a network approach and comparing the results with manually curated datasets, the authors showed that protein families can be discovered in an unsupervised way, without the need to use any *a priori* human expert knowledge [155].

Another network is based on Gene Ontology annotation similarity: the nodes are proteins and the edges between them are drawn based on a certain degree of overlap between their Gene Ontology annotations. In a recent paper [156] such a network is compiled for yeast.

## 8 Concluding remarks

Progress in biology will most certainly require thinking about biological systems as complex networks. We reviewed recent literature on experimental procedures to obtain network models for proteomics, computational approaches to improve their accuracy and how tools from Complex Network Analysis can be used to gain insight in the large-scale organization of such networks.

In particular, we highlighted relationships between network topology and robustness of biological systems. Scale-

free networks were demonstrated to be robust towards random perturbations [99], and indeed protein networks fall in this class. Furthermore, there are links between network measures and phenotypic characteristics, such as the observation that knocking out high-degree proteins in general has more severe impact on lethality than knocking out low-degree proteins. We reported that proteins with related functions tend to be connected, giving an instrument to predict functions of unknown proteins [106].

Every network has its own specific issues, both biological and procedural: which physical mechanisms do the edges represent, what is the meaning of undirected *versus* directed edges, what statistical assumptions have been made to relate nodes, what thresholds have been adopted, *etc?* All these aspects and many more have to be correctly addressed in order to be able to better explore the characteristics of biological systems behavior. Moreover, the quality of network data could heavily influence findings of Complex Network Analysis [84]. Hopefully, future emphasis will be put on constructing high confidence network datasets, by integrating results from different technologies and heterogeneous information sources, in addition to improved experimentation. In the meanwhile, results should be consistently demonstrated on different datasets [102, 103]. Many important findings of Complex Network Analysis depend on comparisons of the protein networks with null-models: if a certain property in the network under consideration is significantly different from what is expected by chance alone, then this property might have interesting biological implications. Selection of the null-model therefore is a crucial step in the analysis of network topologies [33, 157, 158], and care should be taken before making strong biological conclusions.

In future work, PINs and PSNs could be made dynamic, by including information on protein-protein associations that occur or that are lost during a changing environment. Overlapping PINs and PSNs using NetworKIN or protein and peptide chip experiments will enhance the construction of dynamic models of cellular regulation. Additional layers within these networks are provided by other large-scale studies, such as chemical genetics [159], spatio-temporal analysis of promoter activities [160], RNAi and mutant screens, analysis of other PTMs, and by combining them with transcriptional regulatory networks [145]. Monitoring signaling networks on a single-cell level [161] is expected to lead to the formulation of ever more sophisticated network models. Ultimately, networks including all regulatory events occurring in the metab-olome, proteome and transcriptome will become available for Complex Network Analysis. Although this is not expected to happen in the near future, analyzing networks on the level of the proteome will provide many insights into the functional plasticity of organisms.

*We thank the reviewers and the editor for insightful comments and suggestions. SFB is supported by the Austrian Science Foundation, the Vienna Science and Technology Fund and the*

European Union. GM and EC are supported by Sardegna Ricerche. EP and ALF thank Regione Autonoma della Sardegna.

The authors have declared no conflict of interest.

## 9 References

- [1] Oltvai, Z. N., Barabasi, A. L., Systems biology. Life's complexity pyramid. *Science* 2002, **298**, 763–764.
- [2] Barabasi, A. L., Oltvai, Z. N., Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 2004, **5**, 101–113.
- [3] Goh, K. I., Cusick, M. E., Valle, D., Childs, B. *et al.*, The human disease network. *Proc. Natl. Acad. Sci. USA* 2007, **104**, 8685–8690.
- [4] Kann, M. G., Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform.* 2007.
- [5] Newman, M., The structure and function of complex networks. *SIAM Rev.* 2003, **45**, 167–256.
- [6] Albert, R., Scale-free networks in cell biology. *J. Cell. Sci.* 2005, **118**, 4947–4957.
- [7] Park, J., Newman, M. E., Statistical mechanics of networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2004, **70**, 66–117.
- [8] Caldarelli, G., *Scale-free networks*, Oxford University Press 2005.
- [9] Bollobas, B., *Modern Graph Theory*, Springer-Verlag, New York 1998.
- [10] Diestel, R., *Graph Theory*, Springer-Verlag, New York 2000.
- [11] Dorogovtsev, S. N., Mendes, J. F. F., *Evolution of Networks: from biological networks to the Internet and WWW*, Oxford University Press, Oxford 2003.
- [12] Broder, A., Kumar, R., Maghoul, F., Raghavan, P. *et al.*, Graph structure in the Web. *Computer Networks* 2000, **33**, 309–320.
- [13] Zhao, J., Yu, H., Luo, J. H., Cao, Z. W., Li, Y. X., Hierarchical modularity of nested bow-ties in metabolic networks. *BMC Bioinformatics* 2006, **7**, 386.
- [14] Ma, H. W., Zhao, X. M., Yuan, Y. J., Zeng, A. P., Decomposition of metabolic network into functional modules based on the global connectivity structure of reaction graph. *Bioinformatics* 2004, **20**, 1870–1876.
- [15] Ma, H. W., Zeng, A. P., The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* 2003, **19**, 1423–1430.
- [16] Newman, M. E., Mixing patterns in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2003, **67**, 026126.
- [17] Newman, M. E., Assortative mixing in networks. *Phys. Rev. Lett.* 2002, **89**, 208701.
- [18] Watts, D. J., Strogatz, S. H., Collective dynamics of 'small-world' networks. *Nature* 1998, **393**, 440–442.
- [19] Przulj, N., in: Jursica, I., Wigle, D. (Eds.), *Knowledge discovery in proteomics*, CRC Press Boca Raton, FL 2005, pp. 73–146.
- [20] Guelzim, N., Bottani, S., Bourguin, P., Kepes, F., Topological and causal structure of the yeast transcriptional regulatory network. *Nat. Genet.* 2002, **31**, 60–63.
- [21] Mason, O., Verwoerd, M., Graph theory and networks in biology. *IET Syst. Biol.* 2007, **1**, 89–119.
- [22] Erdős, P., Renyi, A., On Random Graphs. *Publ. Math. Debrecen.* 1959, **6**, 290–297.
- [23] Newman, M. E. J., Strogatz, S. H., Watts, D. J., Random graphs with arbitrary degree distributions and their applications. *ArXiv:cond-mat/0007235v2* 2001.
- [24] Albert, R., Barabasi, A. L., Statistical mechanics of complex networks *Rev. Mod. Phys.* 2002, **74**, 47–97.
- [25] Guimera, R., Sales-Pardo, M., Amaral, L. A. N., Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E* 2004, **70**, 025101.
- [26] Barabasi, A. L., Albert, R., Emergence of scaling in random networks. *Science* 1999, **286**, 509–512.
- [27] Ispolatov, I., Krapivsky, P. L., Mazo, I., Yuryev, A., Cliques and duplication-divergence network growth. *New J. Phys.* 2005, **7**, 145.
- [28] Ravasz, E., Barabasi, A. L., Hierarchical organization in complex networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2003, **67**, 026112.
- [29] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N. *et al.*, Network motifs: simple building blocks of complex networks. *Science* 2002, **298**, 824–827.
- [30] Prill, R. J., Iglesias, P. A., Levchenko, A., Dynamic properties of network motifs contribute to biological network organization. *PLoS Biol.* 2005, **3**, e343.
- [31] Christensen, C., Thakar, J., Albert, R., Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET Syst. Biol.* 2007, **1**, 61–77.
- [32] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R. *et al.*, Superfamilies of evolved and designed networks. *Science* 2004, **303**, 1538–1542.
- [33] Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., Stone, L., Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science* 2004, **305**, 1107; author reply 1107.
- [34] Ingram, P. J., Stumpf, M. P., Stark, J., Network motifs: structure does not determine function. *BMC Genomics* 2006, **7**, 108.
- [35] Przulj, N., Biological network comparison using graphlet degree distribution. *Bioinformatics* 2007, **23**, 177–183.
- [36] Przulj, N., Corneil, D. G., Jurisica, I., Modeling interactome: scale-free or geometric? *Bioinformatics* 2004, **20**, 3508–3515.
- [37] Przulj, N., Corneil, D. G., Jurisica, I., Efficient estimation of graphlet frequency distributions in protein-protein interaction networks. *Bioinformatics* 2006, **22**, 974–980.
- [38] Guimera, R., Sales-Pardo, M., Amaral, L. A. N., Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* 2007, **3**, 63–69.
- [39] Maslov, S., Role model for modules. *Nat. Phys.* 2007, **3**, 18–19.
- [40] Girvan, M., Newman, M. E., Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 2002, **99**, 7821–7826.
- [41] Newman, M. E., Fast algorithm for detecting community structure in networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 2004, **69**, 066133.

- [42] Fortunato, S., Barthelemy, M., Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* 2007, *104*, 36–41.
- [43] Arenas, A., Fernandez, A., Gomez, S., Multiple resolution of the modular structure of complex networks. *ArXiv:physics/0703218* 2007.
- [44] Vidal, M., Interactome modeling. *FEBS Lett.* 2005, *579*, 1834–1838.
- [45] Vidal, M., [Network “interactome”]. *Bull. Mem. Acad. R. Med. Belg.* 2006, *161*, 199–210; discussion 210–212.
- [46] Butland, G., Peregrin-Alvarez, J. M., Li, J., Yang, W. *et al.*, Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 2005, *433*, 531–537.
- [47] Ito, T., Chiba, T., Ozawa, R., Yoshida, M. *et al.*, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4569–4574.
- [48] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, *403*, 623–627.
- [49] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A. *et al.*, A protein interaction map of *Drosophila melanogaster*. *Science* 2003, *302*, 1727–1736.
- [50] Uetz, P., Pankratz, M. J., Protein interaction maps on the fly. *Nat. Biotechnol.* 2004, *22*, 43–44.
- [51] Li, S., Armstrong, C. M., Bertin, N., Ge, H. *et al.*, A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, *303*, 540–543.
- [52] Persico, M., Ceol, A., Gavrila, C., Hoffmann, R. *et al.*, HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics* 2005, *6 Suppl 4*, S21.
- [53] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T. *et al.*, Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005, *437*, 1173–1178.
- [54] Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L. *et al.*, Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* 2006, *38*, 285–293.
- [55] Schwikowski, B., Uetz, P., Fields, S., A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 2000, *18*, 1257–1261.
- [56] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P. *et al.*, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002, *30*, 303–305.
- [57] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K. *et al.*, DIP: the database of interacting proteins. *Nucleic Acids Res.* 2000, *28*, 289–291.
- [58] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X. J. *et al.*, DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.* 2001, *29*, 239–241.
- [59] Bader, G. D., Betel, D., Hogue, C. W., BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2003, *31*, 248–250.
- [60] Bader, G. D., Hogue, C. W., BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* 2000, *16*, 465–477.
- [61] Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. *et al.*, BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* 2001, *29*, 242–245.
- [62] Mewes, H. W., Albermann, K., Heumann, K., Liebl, S., Pfeiffer, F., MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res.* 1997, *25*, 28–30.
- [63] Mewes, H. W., Frishman, D., Mayer, K. F., Munsterkotter, M. *et al.*, MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* 2006, *34*, D169–172.
- [64] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G. *et al.*, MINT: a Molecular INTERaction database. *FEBS Lett.* 2002, *513*, 135–140.
- [65] Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G. *et al.*, MINT: the Molecular INTERaction database. *Nucleic Acids Res.* 2007, *35*, D572–D574.
- [66] Joshi-Tope, G., Gillespie, M., Vastrik, I., D’Eustachio, P. *et al.*, Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 2005, *33*, D428–D432.
- [67] Vastrik, I., D’Eustachio, P., Schmidt, E., Joshi-Tope, G. *et al.*, Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 2007, *8*, R39.
- [68] Bork, P., Jensen, L. J., von Mering, C., Ramani, A. K. *et al.*, Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.* 2004, *14*, 292–299.
- [69] Yook, S. H., Oltvai, Z. N., Barabasi, A. L., Functional and topological characterization of protein interaction networks. *Proteomics* 2004, *4*, 928–942.
- [70] Walhout, A. J., Boulton, S. J., Vidal, M., Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast* 2000, *17*, 88–94.
- [71] Grigoriev, A., On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* 2003, *31*, 4157–4161.
- [72] Fields, S., Song, O., A novel genetic system to detect protein-protein interactions. *Nature* 1989, *340*, 245–246.
- [73] Uetz, P., Two-hybrid arrays. *Curr. Opin. Chem. Biol.* 2002, *6*, 57–62.
- [74] Gavin, A. C., Bosche, M., Krause, R., Grandi, P. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, *415*, 141–147.
- [75] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D. *et al.*, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002, *415*, 180–183.
- [76] Krogan, N. J., Cagney, G., Yu, H., Zhong, G. *et al.*, Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, *440*, 637–643.
- [77] Burckstummer, T., Bennett, K. L., Preradovic, A., Schutze, G. *et al.*, An efficient tandem affinity purification procedure for interaction proteomics in mammalian cells. *Nat. Methods* 2006, *3*, 1013–1019.
- [78] Rubio, V., Shen, Y., Saijo, Y., Liu, Y. *et al.*, An alternative tandem affinity purification strategy applied to Arabidopsis protein complex isolation. *Plant J.* 2005, *41*, 767–778.
- [79] Tsai, A., Carstens, R. P., An optimized protocol for protein purification in cultured mammalian cells using a tandem affinity purification approach. *Nat. Protoc.* 2006, *1*, 2820–2827.
- [80] Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F. *et al.*, Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* 2007, *6*, 439–450.

- [81] Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., Wodak, S. J., Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*. *Proteomics* 2007, 7, 944–960.
- [82] Jones, R. B., Gordus, A., Krall, J. A., MacBeath, G., A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* 2006, 439, 168–174.
- [83] Bork, P., Comparative analysis of protein interaction networks. *Bioinformatics* 2002, 18 Suppl 2, S64.
- [84] Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., Vidal, M., Effect of sampling on topology predictions of protein-protein interaction networks. *Nat. Biotechnol.* 2005, 23, 839–844.
- [85] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 2002, 417, 399–403.
- [86] Bader, J. S., Chaudhuri, A., Rothberg, J. M., Chant, J., Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.* 2004, 22, 78–85.
- [87] Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W. *et al.*, Detecting protein function and protein-protein interactions from genome sequences. *Science* 1999, 285, 751–753.
- [88] Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., Eisenberg, D., A combined algorithm for genome-wide prediction of protein function. *Nature* 1999, 402, 83–86.
- [89] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y. *et al.*, A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 2003, 302, 449–453.
- [90] Yamanishi, Y., Vert, J. P., Kanehisa, M., Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics* 2004, 20 Suppl 1, i363–i370.
- [91] von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S. *et al.*, STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 2003, 31, 258–261.
- [92] von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S. *et al.*, STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 2007, 35, D358–D362.
- [93] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D. *et al.*, STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 2005, 33, D433–D437.
- [94] Wagner, A., How the global structure of protein interaction networks evolves. *Proc. Biol. Sc.* 2003, 270, 457–466.
- [95] Wagner, A., The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* 2001, 18, 1283–1292.
- [96] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., Lethality and centrality in protein networks. *Nature* 2001, 411, 41–42.
- [97] Stumpf, M. P., Wiuf, C., May, R. M., Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl. Acad. Sci. USA* 2005, 102, 4221–4224.
- [98] Lin, N., Zhao, H., Are scale-free networks robust to measurement errors? *BMC Bioinformatics* 2005, 6, 119.
- [99] Albert, R., Jeong, H., Barabasi, A. L., Error and attack tolerance of complex networks. *Nature* 2000, 406, 378–382.
- [100] Joy, M. P., Brock, A., Ingber, D. E., Huang, S., High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005, 2005, 96–103.
- [101] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S. *et al.*, Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 2004, 430, 88–93.
- [102] Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L. *et al.*, Still stratus not altocumulus: further evidence against the date/party hub distinction. *PLoS Biol.* 2007, 5, e154.
- [103] Batada, N. N., Reguly, T., Breitkreutz, A., Boucher, L. *et al.*, Stratus not altocumulus: a new view of the yeast protein interaction network. *PLoS Biol.* 2006, 4, e317.
- [104] Maslov, S., Sneppen, K., Protein interaction networks beyond artifacts. *FEBS Lett.* 2002, 530, 255–256.
- [105] Maslov, S., Sneppen, K., Specificity and stability in topology of protein networks. *Science* 2002, 296, 910–913.
- [106] Sharan, R., Ulitsky, I., Shamir, R., Network-based prediction of protein function. *Mol. Syst. Biol.* 2007, 3, 88.
- [107] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T., Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* 2001, 18, 523–531.
- [108] Wuchty, S., Oltvai, Z. N., Barabasi, A. L., Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* 2003, 35, 176–179.
- [109] Remm, M., Storm, C. E., Sonnhammer, E. L., Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 2001, 314, 1041–1052.
- [110] Vespignani, A., Evolution thinks modular. *Nat. Genet.* 2003, 35, 118–119.
- [111] Rives, A. W., Galitski, T., Modular organization of cellular networks. *Proc. Natl. Acad. Sci. USA* 2003, 100, 1128–1133.
- [112] Brohee, S., van Helden, J., Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 2006, 7, 488.
- [113] Van Dongen, S., *Graph Clustering by Flow Simulation*. PhD Thesis, Centers for mathematics and computer science (CWI) University of Utrecht 2000.
- [114] Enright, A. J., Van Dongen, S., Ouzounis, C. A., An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 2002, 30, 1575–1584.
- [115] King, A. D., Przulj, N., Jurisica, I., Protein complex prediction via cost-based clustering. *Bioinformatics* 2004, 20, 3013–3020.
- [116] Blatt, M., Wiseman, S., Domany, E., Superparamagnetic clustering of data. *Phys. Rev. Lett.* 1996, 76, 3251–3254.
- [117] Bader, G. D., Hogue, C. W., An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 2003, 4, 2.
- [118] Frey, B. J., Dueck, D., Clustering by passing messages between data points. *Science* 2007, 315, 972–976.
- [119] Leone, M., Sumedha, Weigt, M., Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics* 2007.
- [120] Seet, B. T., Dikic, I., Zhou, M. M., Pawson, T., Reading protein modifications with interaction domains. *Nat. Rev. Mol. Cell. Biol.* 2006, 7, 473–483.
- [121] Bauch, A., Superti-Furga, G., Charting protein complexes, signaling pathways, and networks in the immune system. *Immunol. Rev.* 2006, 210, 187–207.

- [122] Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P. O. *et al.*, A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nat. Cell Biol.* 2004, 6, 97–105.
- [123] Tewari, M., Hu, P. J., Ahn, J. S., Ayivi-Guedehoussou, N. *et al.*, Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. *Mol. Cell* 2004, 13, 469–482.
- [124] Janes, K. A., Albeck, J. G., Gaudet, S., Sorger, P. K. *et al.*, A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* 2005, 310, 1646–1653.
- [125] Ideker, T., Lauffenburger, D., Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol.* 2003, 21, 255–262.
- [126] Papin, J. A., Hunter, T., Palsson, B. O., Subramaniam, S., Reconstruction of cellular signalling networks and analysis of their properties. *Na. Rev. Mol. Cell. Biol.* 2005, 6, 99–111.
- [127] Huang, P. H., Mukasa, A., Bonavia, R., Flynn, R. A. *et al.*, Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. *Proc. Natl. Acad. Sci. USA* 2007, 104, 12867–12872.
- [128] Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R. *et al.*, Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell* 2004, 15, 3841–3862.
- [129] Chen, K. C., Csikasz-Nagy, A., Gyorffy, B., Val, J. *et al.*, Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* 2000, 11, 369–391.
- [130] Kholodenko, B. N., Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell. Biol.* 2006, 7, 165–176.
- [131] Tyson, J. J., Chen, K., Novak, B., Network dynamics and cell physiology. *Nat. Rev. Mol. Cell. Biol.* 2001, 2, 908–916.
- [132] Pawson, T., Scott, J. D., Protein phosphorylation in signaling—50 years and counting. *Trends Biochem. Sci.* 2005, 30, 286–290.
- [133] de la Fuente, A., Bing, N., Hoeschele, I., Mendes, P., Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 2004, 20, 3565–3574.
- [134] Santos, S. D., Verveer, P. J., Bastiaens, P. I., Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat. Cell. Biol.* 2007, 9, 324–330.
- [135] de la Fuente, A., Brazhnik, P., Mendes, P., Linking the genes: inferring quantitative gene networks from microarray data. *Trends Genet.* 2002, 18, 395–398.
- [136] de la Fuente, A., Brazhnik, P., Mendes, P., in: Yi, T. M., Hucka, M., Morohashi, M., Kitano, H. (Eds.), *2nd Int. Conf. Syst. Biol.*, Omnipress, California Institute of Technology, Pasadena, CA 2001, pp. 213–221.
- [137] Bruggeman, F. J., Kholodenko, B. N., Modular interaction strengths in regulatory networks; an example. *Mol. Biol. Rep.* 2002, 29, 57–61.
- [138] Kholodenko, B. N., Kiyatkin, A., Bruggeman, F. J., Sontag, E. *et al.*, Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proc. Natl. Acad. Sci. USA* 2002, 99, 12841–12846.
- [139] Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P., Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 2005, 308, 523–529.
- [140] Olsen, J. V., Blagoev, B., Gnäd, F., Macek, B. *et al.*, Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* 2006, 127, 635–648.
- [141] Schmelzle, K., White, F. M., Phosphoproteomic approaches to elucidate cellular signaling networks. *Curr. Opin. Biotechnol.* 2006, 17, 406–414.
- [142] Benschop, J. J., Mohammed, S., O'Flaherty, M., Heck, A. J. *et al.*, Quantitative phosphoproteomics of early elicitor signaling in *Arabidopsis*. *Mol. Cell. Proteomics* 2007, 6, 1198–1214.
- [143] Gruhler, A., Olsen, J. V., Mohammed, S., Mortensen, P. *et al.*, Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* 2005, 4, 310–327.
- [144] Munton, R. P., Tweedie-Cullen, R., Livingstone-Zatchej, M., Weinandy, F. *et al.*, Qualitative and quantitative analyses of protein phosphorylation in naive and stimulated mouse synaptosomal preparations. *Mol. Cell. Proteomics* 2007, 6, 283–293.
- [145] Ptacek, J., Devgan, G., Michaud, G., Zhu, H. *et al.*, Global analysis of protein phosphorylation in yeast. *Nature* 2005, 438, 679–684.
- [146] Zhu, H., Klemic, J. F., Chang, S., Bertone, P. *et al.*, Analysis of yeast protein kinases using protein chips. *Nat. Genet.* 2000, 26, 283–289.
- [147] Diks, S. H., Kok, K., O'Toole, T., Hommes, D. W. *et al.*, Kinome profiling for studying lipopolysaccharide signal transduction in human peripheral blood mononuclear cells. *J. Biol. Chem.* 2004, 279, 49206–49213.
- [148] Lemeer, S., Jopling, C., Naji, F., Ruijtenbeek, R. *et al.*, Protein-tyrosine kinase activity profiling in knock down zebrafish embryos. *PLoS ONE* 2007, 2, e581.
- [149] de la Fuente van Bentem, S., Hirt, H., Using phosphoproteomics to reveal signalling dynamics in plants. *Trends Plant Sci.* 2007.
- [150] de la Fuente van Bentem, S., Roitinger, E., Anrather, D., Csaszar, E., Hirt, H., Phosphoproteomics as a tool to unravel plant regulatory mechanisms. *Physiologia Plantarum* 2006, 126, 110–119.
- [151] Hutti, J. E., Jarrell, E. T., Chang, J. D., Abbott, D. W. *et al.*, A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods* 2004, 1, 27–29.
- [152] Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. *et al.*, Systematic discovery of *in vivo* phosphorylation networks. *Cell* 2007.
- [153] Bhattacharyya, R. P., Remenyi, A., Good, M. C., Bashor, C. J. *et al.*, The Ste5 scaffold allosterically modulates signaling output of the yeast mating pathway. *Science* 2006, 311, 822–826.
- [154] Ptacek, J., Snyder, M., Charging it up: global analysis of protein phosphorylation. *Trends Genet.* 2006, 22, 545–554.
- [155] Medini, D., Covacci, A., Donati, C., Protein homology network families reveal step-wise diversification of type III and type IV secretion systems. *PLoS Comput Biol* 2006, 2, e173.
- [156] Wu, X., Zhu, L., Guo, J., Fu, C. *et al.*, SPIDER: *Saccharomyces* protein-protein interaction database. *BMC Bioinformatics* 2006, 7 Suppl 5, S16.

- [157] Amaral, L. A. N., Guimera, R., Lies, damned lies and statistics. *Nat. Phys.* 2006, 2, 75–76.
- [158] Colizza, V., Flammini, A., Serrano, M. A., Vespignani, V., Detecting rich-club ordering in complex networks. *Nat. Phys.* 2006, 2, 110–115.
- [159] Knight, Z. A., Shokat, K. M., Features of selective kinase inhibitors. *Chem. Biol.* 2005, 12, 621–637.
- [160] Dupuy, D., Bertin, N., Hidalgo, C. A., Venkatesan, K. *et al.*, Genome-scale analysis of *in vivo* spatiotemporal promoter activity in *Caenorhabditis elegans*. *Nat. Biotechnol.* 2007, 25, 663–668.
- [161] Sachs, K., Gifford, D., Jaakkola, T., Sorger, P., Lauffenburger, D. A., Bayesian network approach to cell signaling pathway modeling. *Sci. STKE* 2002, 2002, PE38.
- [162] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003, 13, 2498–2504.