Vella *et al. EURASIP Journal on Bioinformatics and Systems Biology* (2017) 2017:6
DOI 10.1186/s13637-017-0059-z

EURASIP Journal on Bioinformatics and Systems Biology

REVIEW Open Access

From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data

Danila Vella[1,2], Italo Zoppis[2], Giancarlo Mauri[2], Pierluigi Mauri[1] and Dario Di Silvestre[1*]

Abstract

The reductionist approach of dissecting biological systems into their constituents has been successful in the first stage of the molecular biology to elucidate the chemical basis of several biological processes. This knowledge helped biologists to understand the complexity of the biological systems evidencing that most biological functions do not arise from individual molecules; thus, realizing that the emergent properties of the biological systems cannot be explained or be predicted by investigating individual molecules without taking into consideration their relations. Thanks to the improvement of the current -omics technologies and the increasing understanding of the molecular relationships, even more studies are evaluating the biological systems through approaches based on graph theory. Genomic and proteomic data are often combined with protein-protein interaction (PPI) networks whose structure is routinely analyzed by algorithms and tools to characterize hubs/bottlenecks and topological, functional, and disease modules. On the other hand, co-expression networks represent a complementary procedure that give the opportunity to evaluate at system level including organisms that lack information on PPIs. Based on these premises, we introduce the reader to the PPI and to the co-expression networks, including aspects of reconstruction and analysis. In particular, the new idea to evaluate large-scale proteomic data by means of co-expression networks will be discussed presenting some examples of application. Their use to infer biological knowledge will be shown, and a special attention will be devoted to the topological and module analysis.

Keywords: Co-expression network, -Omics data, PPI network, Systems biology, Topological analysis, WGCNA, Pearson's correlation

1 Introduction

The development of systems biology approaches based on graph theory [1–3] is receiving a great boost by the improvement of the -omics technologies that allow more and more big amount of accurate qualitative and quantitative measures [4, 5]. New methodologies have also been developed to increase knowledge about protein-protein interactions (PPIs) [6]. As a result, the PPI networks combined with protein and with gene expression levels are today widespread to investigate biological systems [7–10].

The magnitude of -omics data provides the opportunity to decode in alternative way the role of biological molecules and processes characterizing the emergent phenotypes. In this scenario, a common procedure to evaluate gene expression levels is based on statistics that measure the dependence between variables, and the resulting co-expression networks are used to identify genes functionally related or controlled by the same transcriptional regulatory program [11–13]. Unlike gene expression levels, the use of proteomic data to infer co-expression networks has been explored through few studies [14–20]. Similar to PPI and gene co-expression networks, these networks have been evaluated at topological level in terms of edge rearrangement, as well as of modules associated with common cellular functions.

*Correspondence: dario.disilvestre@itb.cnr.it
[1] Institute for Biomedical Technologies - National Research Council (ITB-CNR), 93 Fratelli Cervi, Segrate, Milan, Italy
Full list of author information is available at the end of the article

Although different aspects including data collection and network reconstruction need to be improved, the preliminary results are proving this approach promising as alternative to evaluate large-scale proteomic data. This could have important effects into clinical applications favoring the way toward the use of multiple biomarkers and their relationships [17, 21–24]. Thus, in addition to improve basic research, these elements may contribute to develop most efficient diagnosis and prognosis methods to a more preventive, predictive, and personalized medicine [25–27].

Based on these premises, in this review we introduce the reader to PPI and co-expression networks. The recent idea to describe and to evaluate proteomic data by means of co-expression networks will be discussed presenting some example of application. Their use will be shown to infer biological knowledge, and a special attention will be devoted to the topological and module analysis.

2 Protein interaction networks

Graph theory is a powerful abstracting machinery that allows to model several types of system, both natural and human-made, ranging from biology to sociology science [28]. A graph, also called network, provides a system representation in terms of relationships among the elements that make it up; a set of nodes V, stands for the elements of the system, while a set of edges E, stands for their relations. Mathematically, we refer to a graph as $G = (V, E)$ (Fig. 1a).

Concerning biological networks, the nodes may be correlated of attributes representing characteristics of interest, such as expression levels or GO terms. In the same way, the edges may possess attributes describing the relation between nodes, for example indicating the strength of the interaction or its reliability; edges may also be directed or undirected, and here we shall mainly deal with undirected edges. Using the framework described in Fig. 1, a protein interaction network is defined as a complex graph, where the nodes are proteins and the edges represent their relation, generally physical or functional, like proposed by Vidal et al. [25].

2.1 PPI: physical and functional protein links

A protein interaction network usually refers to physical PPIs [29], but several meanings have been attributed to this term. In fact, a group of proteins working together to perform a biological function not necessarily are in direct contact, but their relation may be of regulation or influence, for example, making use of intermediary molecules. For this reason, the term PPI has not only been

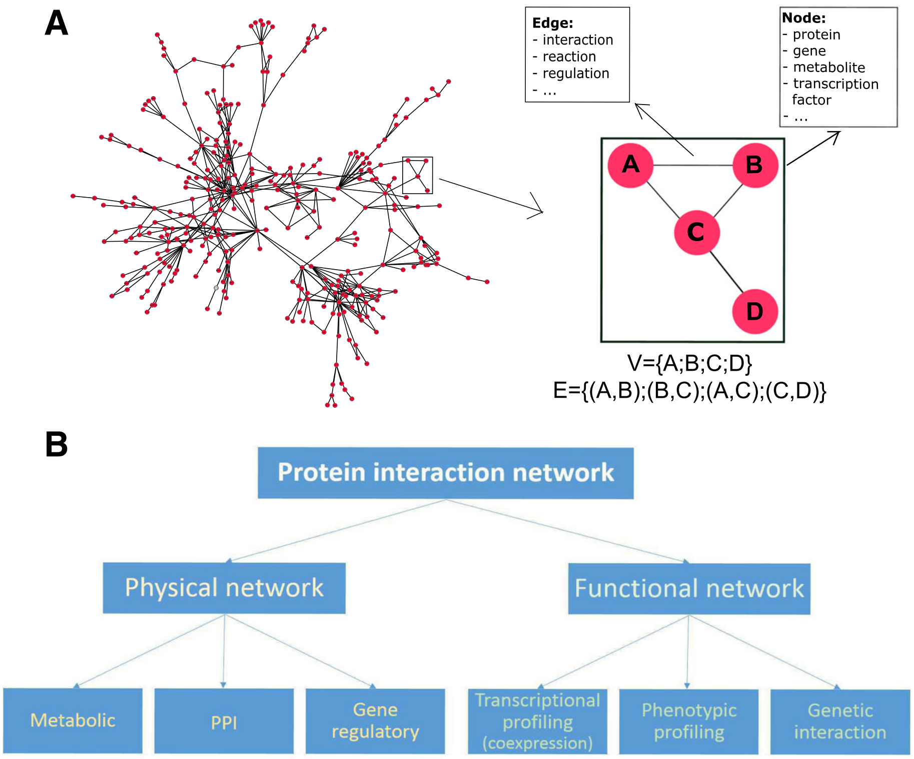


Fig. 1 **a** Biological networks. Nodes may represent several types of biological elements, while the edges describe the nature of their relationship. If *A* and *B* are two nodes connected by an edge, $(A, B) \in E$, *B* is a *neighbor* of *A* or *A* and *B* are adjacent. **b** Protein network classification proposed by Vidal et al. [25]

exclusively used to indicate a physical contact between proteins, but also proteins connected by functional links. It is important to bear in mind that proteins participate to physical-chemical connection depending on the biological context where they are [30]. Thus, the interactions composing a given network could not occur in any cell or at any time. However, if two interacting proteins are experimentally identified in a given sample, we assume they also interact in the system we are studying, thus their relation is reported in the reconstructed PPI network to be analyzed.

2.2 PPI: detection, storage, and analysis tools

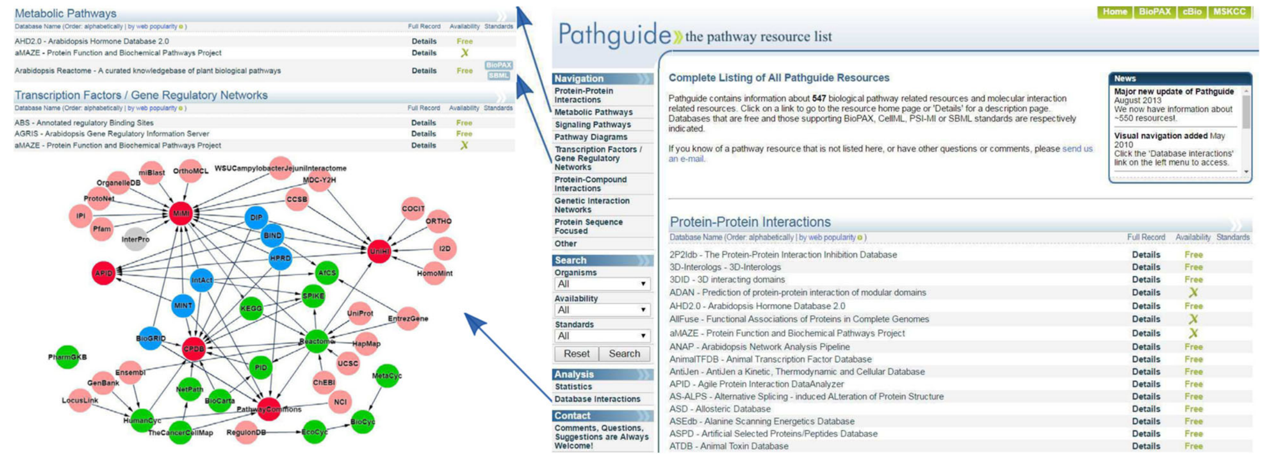
The main approaches to demonstrate physical interaction between proteins are the yeast two-hybrid (Y2H) method and the tandem affinity purification coupled with mass-spectrometry (TAP-MS) [6]. To reduce the identification of false interactions, these experimental data are complemented with computational methods of prediction [31–33]. Other methods are used to identify functional relationships, and most of them rely on protein expression data [20], analysis of gene co-expression patterns [34], and analysis of sequences or phylogenetic properties, as Rosetta Stone or Sequence co-evolution methods [35].

Both physical and functional PPIs are stored in public repositories. The most popular include MINT [36], IntAct [37], STRING [38], and HPRD [39]. The latter specifically collects interactions related to *Homo sapiens*, while other databases like STRING collect different kinds of interactions (from experiments/biochemistry, annotated pathways, gene neighborhood, gene fusion, gene co-occurrence, gene co-expression, and text-mining) and different organisms. A useful list of repositories presented by De Las Rivas et al. [29] provides a classification in categories (primary, meta, and prediction database) according to method used to detect interactions. Moreover, an exhaustive collection of more than 500 databases is available in the Pathguide website (Fig. 2) [40].

The development of computational tools to retrieve, visualize, and analyze biological networks is a key aspect of the systems biology studies, like the production of accurate -omics data and the collection of reliable molecular interactions. The most broadly adopted softwares include Cytoscape and its plugins [41], VisANT [42], atBioNet [43], PINA [44], and Ingenuity [45] which represents a commercial solution. On the contrary, Cytoscape is a software now developed by an international consortium of open-source developers. Figure 3 shows a possible use of the ReactomeFIViz Cytoscape's plugin to obtain networks (both functional and physical) associated with a given biological function. ReactomeFIViz is focused to pathways and patterns related to cancer and other pathologies [46]. This is of importance in the context of biomedical research, and detailed reviews about network models to investigate complex diseases have been published by Cho et al. [47] and by Vidal et al. [25]. Both works show how functional and physical links can be used to investigate disease mechanisms, and PPI networks emerge as effective model to evaluate different biomolecules acting in complex biological systems, thus providing an insight on phenomenons involved in a given physio-pathological context.

3 Co-expression networks

The great amount of data produced by microarray and RNA-seq technologies has driven the need of methods to objectively extract meaningful informations, such as genes differentially expressed or sharing a similar expression pattern. A widely adopted approach to evaluate transcript levels is based on statistics that measure the dependence between variables [48]. Co-expression represents the first step of inference that defines a relation between pairs



Metabolic Pathways

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
AHD2.0 - Arabidopsis Hormone Database 2.0	Details	Free	
aMAZE - Protein Function and Biochemical Pathways Project	Details	X	
Arabidopsis Reactome - A curated knowledgebase of plant biological pathways	Details	Free	BioPAX SBML

Transcription Factors / Gene Regulatory Networks

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
ABS - Annotated regulatory Binding Sites	Details	Free	
AGRIS - Arabidopsis Gene Regulatory Information Server	Details	Free	
aMAZE - Protein Function and Biochemical Pathways Project	Details	X	

Home BioPAX cBio MSKCC

Pathguide » the pathway resource list

Navigation
- Protein-Protein Interactions
- Metabolic Pathways
- Signaling Pathways
- Pathway Diagrams
- Transcription Factors / Gene Regulatory Networks
- Protein-Compound Interactions
- Genetic Interaction Networks
- Protein Sequence Focused
- Other

Search

Organisms: All ▾

Availability: All ▾

Standards: All ▾

Reset | Search

Analysis
- Statistics
- Database Interactions

Contact
- Comments, Questions, Suggestions are Always Welcome!

Complete Listing of All Pathguide Resources

Pathguide contains information about **547** biological pathway related resources and molecular interaction related resources. Click on a link to go to the resource home page or 'Details' for a description page. Databases that are free and those supporting BioPAX, CellML, PSI-MI or SBML standards are respectively indicated.

If you know of a pathway resource that is not listed here, or have other questions or comments, please send us an e-mail.

News

Major new update of Pathguide August 2013
We now have information about ~550 resources!.

Visual navigation added May 2010
Click the 'Database interactions' link on the left menu to access.

Protein-Protein Interactions

Database Name (Order: alphabetically | by web popularity)	Full Record	Availability	Standards
2P2Idb - The Protein-Protein Interaction Inhibition Database	Details	Free	
3D-Interologs - 3D-Interologs	Details	Free	
3DID - 3D interacting domains	Details	Free	
ADAN - Prediction of protein-protein interaction of modular domains	Details	X	
AHD2.0 - Arabidopsis Hormone Database 2.0	Details	Free	
AllFuse - Functional Associations of Proteins in Complete Genomes	Details	X	
aMAZE - Protein Function and Biochemical Pathways Project	Details	X	
ANAP - Arabidopsis Network Analysis Pipeline	Details	Free	
AnimalTFDB - Animal Transcription Factor Database	Details	Free	
AntiJen - AntiJen a Kinetic, Thermodynamic and Cellular Database	Details	Free	
APID - Agile Protein Interaction DataAnalyzer	Details	Free	
AS-ALPS - Alternative Splicing - induced ALteration of Protein Structure	Details	Free	
ASD - Allosteric Database	Details	Free	
ASEdb - Alanine Scanning Energetics Database	Details	Free	
ASPD - Artificial Selected Proteins/Peptides Database	Details	Free	
ATDB - Animal Toxin Database	Details	Free	

Fig. 2 Pathguide website [40]. A repository containing information about 547 resources of molecular interactions and pathways

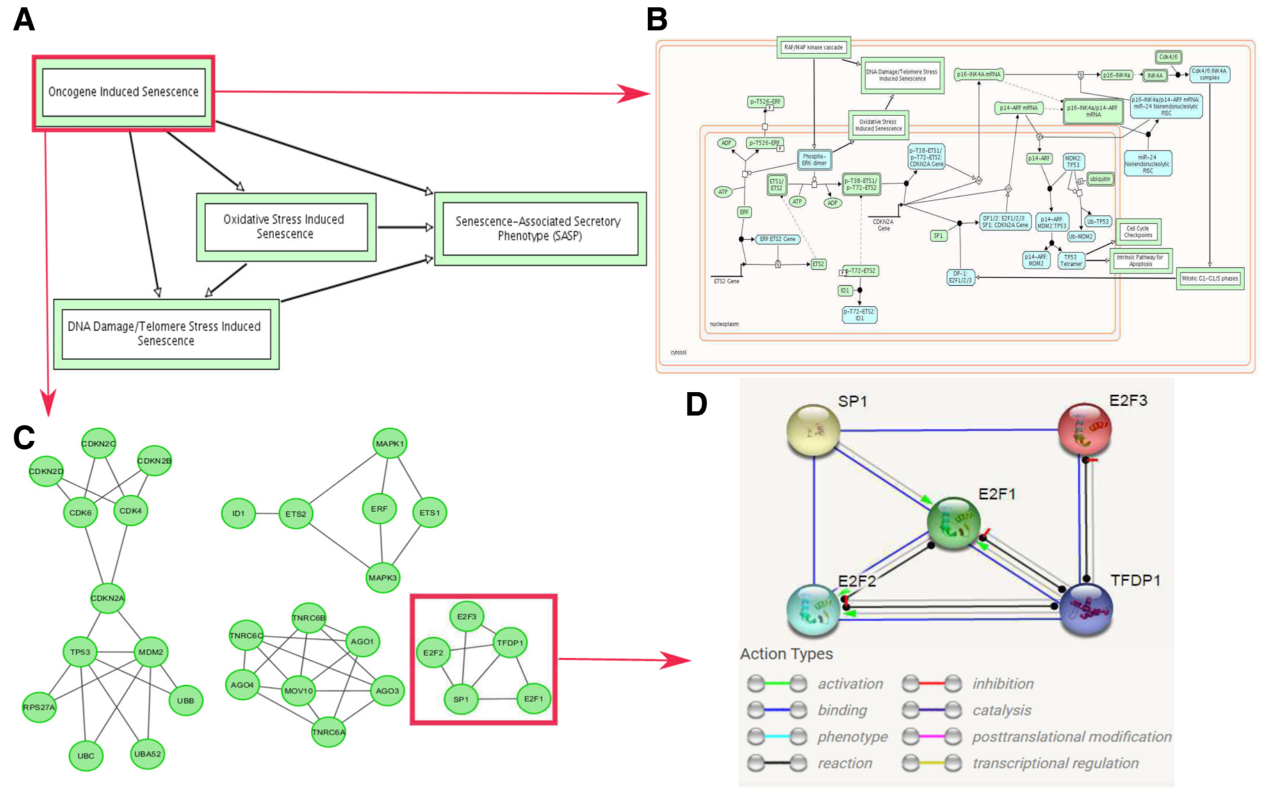


Fig. 3 ReactomeFIViz: from disease pathway to PPI network. Main steps to obtain a protein functional and a physical protein network, starting from a specific pathway (oncogene induced senescence). Using ReactomeFIViz, pathways can be visualized in relation with others (**a**), can be detailed as a diagram showing all intermolecular relationships (**b**), and as a protein functional interaction network (**c**) showing just the relation among proteins that cooperate to perform a given molecular function. Finally, starting from a group of protein of interest, it is possible to obtain a network of protein-protein interactions by STRING; in the reported example, the interactions shown are limited to physical type, in particular binding, activation and inhibition (**d**)

of transcripts. It is based on the concept that transcript profiles of time series, or result of specific perturbations, may be indicative of dynamics and differences between transcripts, implying their regulation. Following the processing of transcript levels, the result is a co-expression network defined as an undirected graphs where the nodes correspond to genes, and the edges indicate significant co-expression relationships, but not causality. This aspect is faced in the context of transcriptional regulatory networks [49], where pairs of genes are considered in a systemic perspective of cooperation, including co-regulation, activation/suppression, and indirect control through the action of siRNA, miRNA, proteins, metabolites, and epigenetic mechanisms. This complexity make difficult the inference of transcriptional regulatory networks by using exclusively transcriptional profiles. In fact, in addition to co-expression, next levels of inference require more information and different modeling techniques, including Boolean networks, Bayesian networks, or differential equations (ODEs), which are revised in more detail in studies addressing *reverse engineering* approaches [49].

Gene co-expression networks are topologically analyzed to identify hubs/bottlenecks and node communities sharing high co-expression score; communities are the starting point to identify topological, functional, and/or disease modules related to specific biological phenotypes [50, 51]. Different studies have shown that genes functionally related, and sharing Gene Ontology (GO) terms, usually present higher co-expression score [52]. Moreover, variations of the co-expression score are evaluated to select topological relevant nodes whose number of interactions changes under specific conditions or perturbations [18] (Fig. 4).

In the last 10 years, the improvement of the liquid chromatography and the mass spectrometry has given a great boost to large-scale proteomics analysis, making available the expression profiles of thousands of proteins per sample [53]. Due to the similarity between gene and protein matrices, the use of proteomic data to infer protein co-expression networks has been recently explored to investigate the role of proteins in specific physio-pathological contexts. Although different aspects need to be improved,

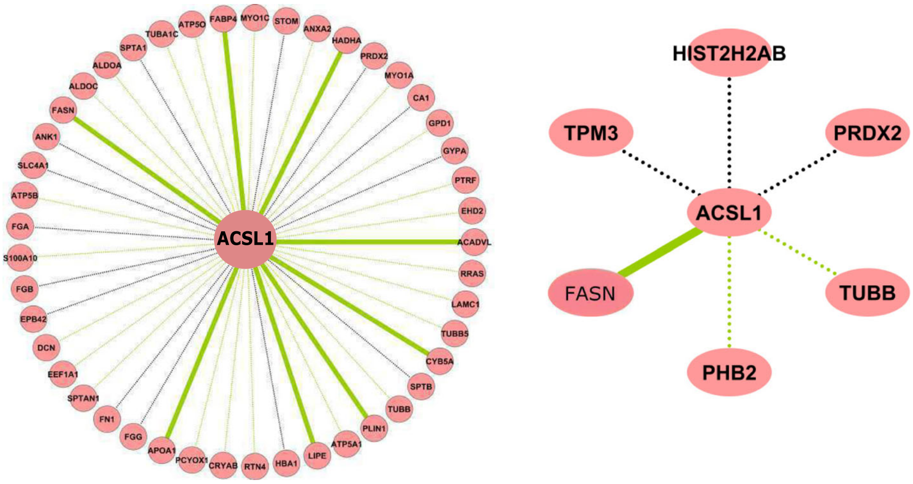


Fig. 4 The figure shows the ACSL1 protein and its neighbors in two co-expression networks obtained by processing the protein expression profiles of a control group and a group of patients affected by amyloidosis disease. In the considered groups of samples, ACSL1 shows a different degree. It suggests that this protein may have a key role in the emergent phanotypes. *Green edges* represent a positive correlation between the expression profiles, while *black edges* indicate negative correlations. The *thick edges* indicate known interactions present in public repositories as PPI

this approach takes into account protein relationships, and, with respect to conventional methods, it represents an alternative to gain a deeper insight of the protein characterizing a given system. This issue will be discussed with greater detail in the paragraph 5.

3.1 Aspects of construction

To build a co-expression network, an important aspect concerns the computation of a co-expression score, which weigh the correlation of two genes/proteins in response to the considered conditions (Fig. 5). To address this issue, metrics to measure gene/protein co-expression have to be considered (Table 1); the most used metrics include Pearson's correlation (PC), Spearman's correlation, Kendall's correlation, and mutual information [48, 54]. Various methods have been also proposed to define proper thresholds to select significant relations. Some of them are based on statistical analysis [55] and on network properties [56], while other interesting approaches aim to minimize the false positive links [57]. Finally, not less important is the selection of appropriate experimental samples/conditions to be processed. A condition-independent analysis is used to find relations of co-expression actual in different biological contexts; on the contrary, a condition-dependent analysis aims to find relations associated with specific phenotypes.

The co-expression score computation may be faced by using any statistical or computational tool that allows to evaluate the dependence between variables. Some tools have been specifically designed to construct, visualize, and analyze co-expression networks. For example, the ExpressionCorrelation Cytoscape's plugin allows to process microarray data and provides a similarity matrix computed by PC [58]. In addition to being user-friendly, the main advantage of this tool is that the reconstructed networks are directly imported in Cytoscape where it may be evaluated by other plugins.

WGCNA is one of the most used approaches to build and to analyze gene co-expression networks [59], and it has been recently adapted for proteomics use also [14–20].

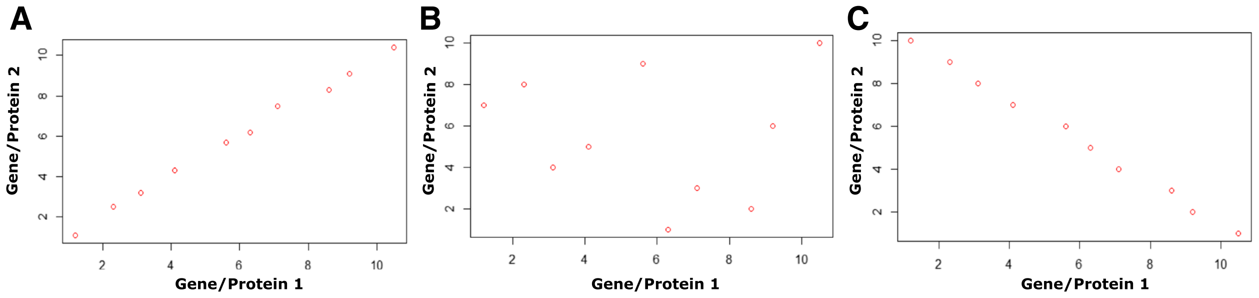


Fig. 5 Possible cases of correlation between two variables. **a** Positive correlation. **b** No correlation. **c** Negative correlation

Table 1 Measures of dependence between two variables

Co-expression measures	What measures?	Input/Output	Features
Pearson's correlation (PC)	Tendency to respond in opposite/same direction across different samples	Input: gene expressions value Output: • $[0, 1]$ both genes increase • $[-1, 0]$ one increase and other decrease	• Sensitivity to outliers • Bad array of expression level can determine positive PC value • Measure linear relations
Spearman's correlation (SC)	Tendency to respond in opposite/same direction across different samples	Input: ranking values from expression levels in samples Output: • $[0, 1]$ Both genes increase • $[-1, 0]$ One increase and the other decrease	• Robust to outliers • Detect non-linear associations
Mutual information	Reduction of uncertainty of a gene given the knowledge about other gene	Input: gene expression values Output: • 0 there is no interdependence • >0 there is interdependence	• Measure complex non-linear type relations (rarely present in biological data) • More samples are needed than PC, SC • Time-consuming computation
Kendall	Correspondence/compatibility among two rankings	Input: gene expression value Output: • 1 perfect correspondence • -1 rankings exactly inverted	• Similar to SC • Robust to outliers • Assumes fewer values than SC in the range $[-1, 1]$

It provides a weighted network model by converting a co-expression measure to a connection weight. The network is fully specified by an adjacency matrix, where the component a_{ij} defines the strength of connection between nodes i and j. The value of a_{ij} is computed through the co-expression similarity s_{ij} (1), defined as the absolute value of correlation among the profiles of nodes i and j. It can be defined in two ways: to obtain an unweighted network, the s_{ij} is filtered by a threshold τ such that a_{ij} takes on value [0,1] (hard-thresholding) (2), while to obtain a weighted network a_{ij} is defined by a power adjacency function (soft-thresholding) (3):

$$s_{\text{ij}} = |\text{cor}(i,j)| \tag{1}$$

$$a_{\text{ij}} = \begin{cases} 1 & s_{\text{ij}} \geq \tau \\ 0 & s_{\text{ij}} < \tau \end{cases} \tag{2}$$

$$a_{\text{ij}} = s_{\text{ij}}^{\beta} \tag{3}$$

The R WGCNA package provides the possibility to use different types of metrics, including Spearman', Pearson', Kendall's correlation (see function *cor*), and the biweight midcorrelation (see function *bicor*) [60]. Spearman's correlation is a non-parametric measure of correlation. Pearson's correlation can be used when data are normally distributed, but it is quite susceptible to the presence of outliers. In this case, the biweight midcorrelation is recommended because it is more robust to outliers. The package allows to compute both the correlation and the Student p value for multiple correlations in case of missing data (see function *corAndPvalue* and *bicorAndPvalue*), while the function *qvalue* computes the q value to measure the significance of each feature in terms of false discovery rate rather than false positive rate [61]. The unweighted network displays sensitivity to the choice of the correlation values cut-off, thus, it is important to use a proper criterion to select the edges to include in the network. It is important to take into account the correlations are computed among each pairs of genes/proteins leading to a high rate of false positive values. Thus, to build an unweighted network and to reduce the inclusion of not significant correlations, it is recommended to set a cut-off also for p and q values. Concerning the weighted networks, the choice of the β parameter is based on the scale-free topology criterion [62]. This method represents an improvement over unweighted networks based on dichotomizing the correlation matrix; the continuous nature of the gene co-expression information is preserved, and the results of weighted network analyses are highly robust with respect to the choice of the parameter β (soft-thresholding power). However, this thresholding method is based on the assumption that the network follows a scale-free topology, a hypothesis weak in some cases, as discussed in Section 4.1.

3.2 WGCNA and proteomic issues

When the WGCNA is applied to proteomic or to metabolomic data, the choice of the optimal cutting parameters should be evaluated in relation to the nature of the data analyzed. In fact, due to the low coverage of the current analytical technologies, the produced dataset are often incomplete, and the methods need to be properly modified [63]. A major concern is the high rate of missing values that introduce loss of information and significant bias. To address this issue, several approaches including K nearest neighbor, least square methods, or local least square methods have been proposed for proteomic and metabolomic datasets too [64]. In other cases, a very simple approach has been adopted, such as the removal of all species with a number of missing data bigger than a given threshold [65]. However, to implement a more accurate analysis, it is recommended to process data by using an imputation method taking into account the nature of missing data. Three types of missing value have been identified: MCAR (missing completely at random), i.e., due to stochastic fluctuations in a proteomic dataset, MAR (missing at a random), i.e., due to multiple minor errors, and MNAR (missing not at a random), i.e., due to limits of abundance of peptides/proteins that instruments are able to detect. In general, methods work fine when a low percentage of missing value ($\leq$10%) is present, but this threshold could be different in relation to the missingness mechanisms and imputation approach used [63, 64].

In addition to missing value, another important step of proteomic data preprocessing concerns their normalization [66]. Batch effects may occur in datasets run in different days or by different technicians. This phenomenon may increase by using isotope reagents which allow the quantitation of a limited number of samples, thus, preventing a simultaneous analysis of multiple samples which could reduce data heterogeneity. For these reasons, an appropriate data transformation is a prerequisite to capture true correlations. Also in the case of protein co-expression, valid correlations have to be selected by applying proper thresholds. To date, the most applications of WGCNA method on proteomic datasets used a the soft-thresholding, which defines the β value according the scale-free criterion [15, 16, 65]. However, since the application of WGCNA to proteomic dataset is a recent issue, and literature reports, few examples, the future evaluation of hard-thresholding approach might be useful.

4 Network topological analysis

The structure of biological networks is closely related to the biological functions performed by a system (cell or tissue) under a given condition. Starting from this point, many studies aim to face biological questions by investigating the network models in terms of topology [67] and modular properties [68]. As for theoretical mathematical models proposed to describe the biological networks, the most claimed are Erdös–Rényi random graphs [69] and scale-free [70] (see Fig. 6). Other models, such as the geometric random graph (GEO) [71] and the small-world [72], have recently been proposed. In the context of biology, the random graph, proposed in 1950, has been overtaken by the scale-free model; in fact, the degree distribution of the scale-free model is a power-law curve that fits better than Poisson curve (typical of random graphs) the degree distribution of the experimental networks [70] (Fig. 7). Based on power law distribution, most nodes have a degree value far from the mean; specifically, most nodes have a low number of interactions while few nodes have a high number of interactions. These features lead a network structure less vulnerable and make the related system biologically robust [73]. Of note, the degree distribution may reflect the different role of proteins/genes, and those with a highest number of connections, so-called hubs, have a higher probability to be more biologically relevant than others. In other words, removal or modification of hubs may induce stronger alteration of the system equilibrium rather than removal or modification of nodes with low degree [74].

Although some topological properties are well described by a theoretical model, it may not be enough to affirm that the model represents well the real-world network considered [75]. For example, a study on PPI network of *Drosophila Melanogaster* and *Saccharomyces Cerevisiae* showed that the degree distribution was in agreement with scale-free model, but diameter, cluster coefficient, and graphlet frequency were closer to GEO [76]. Of note, based on graphlet frequency, the comparison among scale-free, random graph, and GEO models has shown a higher agreement of GEO with PPI network from eukaryotic organism [77, 78]. A possible reason of these findings is that the scale-free model fits networks that emerged from a stochastic growth, not subjected to an optimization process; while, PPI networks emerge from stochastic processes, and their structure is influenced by the evolutionary optimization that living systems have gone through [76].

Another model used to describe the PPI networks is the small-world. Like the random graph model, it is characterized by a Poisson curve. In a study focused on the investigation of proteins regulating the fat storage, the corresponding PPI network had a degree-distribution close to a Poisson curve rather than a power-law [79]. Moreover, the network showed a low average path length and a high clustering coefficient typical of small-world model. These parameters indicate a network organized

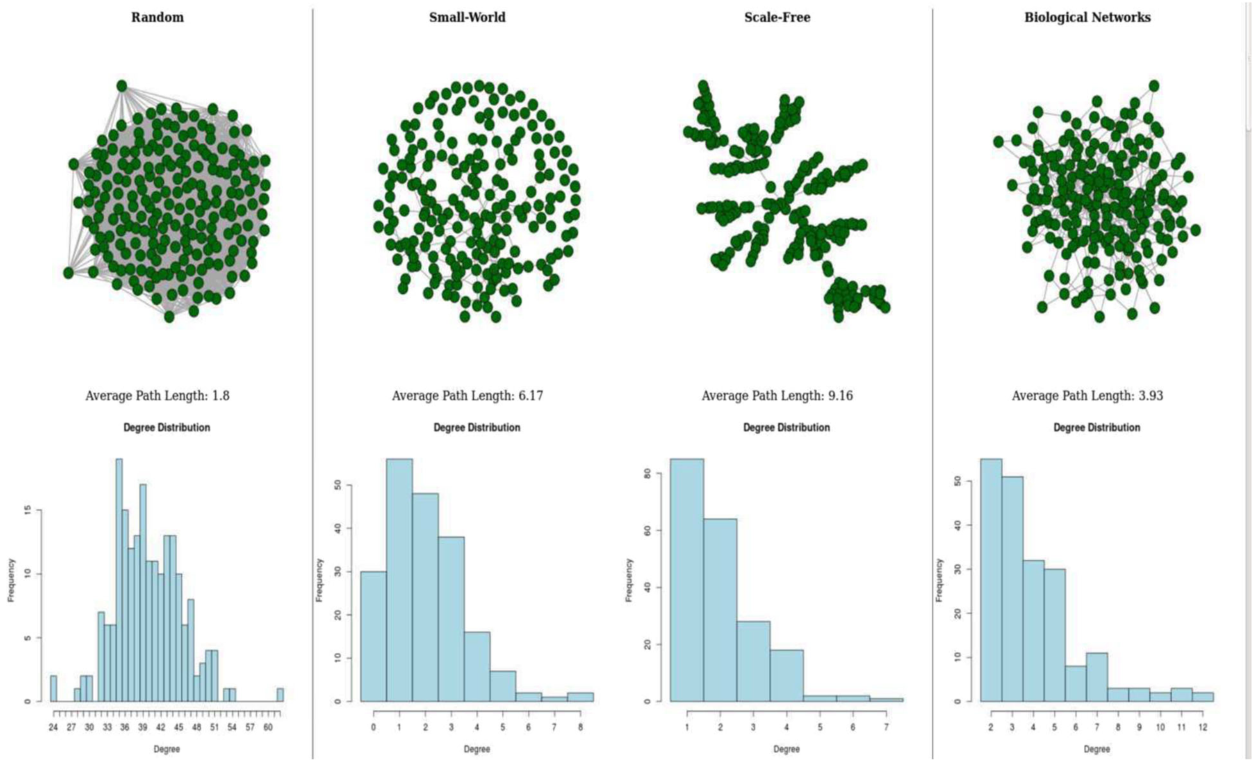


Fig. 6 Shape and degree distribution of random, small-world, and scale-free model with respect to a biological network. Models were calculated by ELIXIR web tool [131]

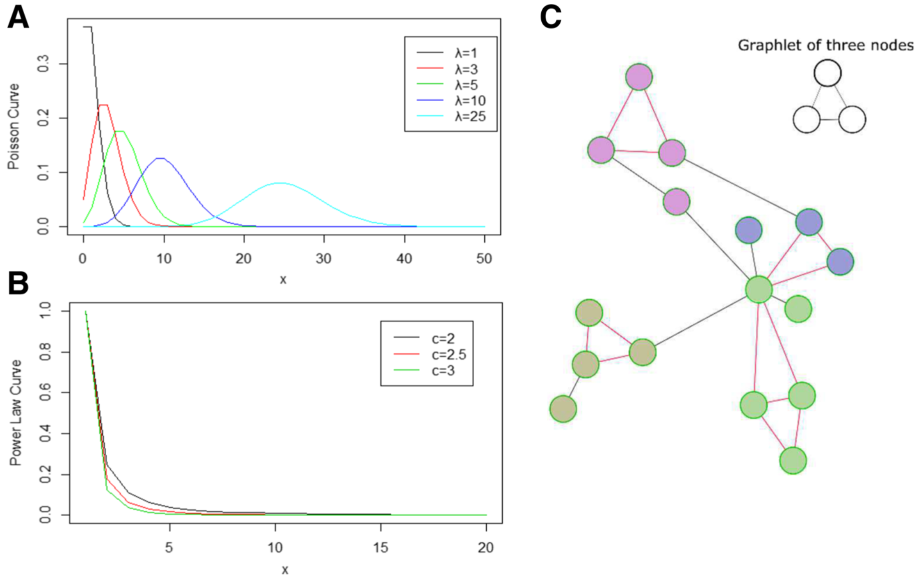


Fig. 7 Functions used to describe the degree distribution of biological networks. Poisson curve **a** and power-law **b** shown for different parameters. **c** Example of graphlet of three nodes with frequency equal to 5

into communities, like observed in PPI networks [80]. The small-world model preserves a modularity structure, and it is not characterized by hub nodes making the small-world networks more robust in the case of removal or modification of any node [73].

The topological evaluation of gene co-expression networks has shown that they are characterized by small-world and by scale-free properties, similar to many other real-world networks [81]. A study showed that the co-expression networks generated from large datasets are scale-free, but with an average clustering coefficient of several orders of magnitude higher than expected for similarly sized scale-free networks [82]. These opposite findings could be explained by the evidence that the topological properties of the co-expression networks may be influenced by different parameters, including the expression data or the similarity measures to evaluate the dependency between variables.

4.1 Topological analysis

A key point of topological studies is the definition of mathematical models and metrics to describe the network's properties and to select the most relevant nodes and substructures that may be of biological significance. Generally called centralities, metrics can be subdivided into measures related to nodes, edges, or whole network. Table 2 lists the main basic centralities used in the network topological analysis [83].

In the context of network organization, these centralities facilitate the answer to question about which proteins are most important and why. To give an idea of such analysis, we say that a vertex (i.e., a protein) is important (or central) if it is close to many other vertexes. There are many number of different centrality measures that have been proposed in literature but probably the most applied, and simple, is called *vertex degree*. The degree $d(v)$ of a vertex v, in a network $G = (V, E)$, counts the number of edges in E incident upon v. Given G, define $f(d)$ to be the fraction of vertexes $v \in V$ with degree $d(v) = d$. For different $d_1, d_2, \ldots, d_n$, the collection $\{f(d_1), f(d_2), \ldots, f(d_n)\}$ is called the degree distribution of G.

A useful generalization of degree is the notion of vertex strength, which is obtained simply by summing up the weights of edges incident to a given vertex. The distribution of strength is sometimes called the weighted degree distributions defined in analogy to the ordinary degree distribution.

Table 2 Centralities calculated by the CentiScaPe Cytoscape's plugin

Centrality	Description	Biological meaning
Diameter[a]	Defines the longest shortest path in the network	
Average distance[a]	Defines the mean length of all the shortest paths in the network	
Degree[b]	Describes the number of neighbors a node has	Highlights the number of nodes that regulated/regulate the node *v*
Eccentricity[b]	Describes the longest shortest paths a node develop, giving us a proximity information	Highlights the easiness of a protein to reach/to be reached by all the other proteins in the network
Closeness[b]	Describes, for the node *v*, the minimal sum of all the distances in the network	Highlights the probability of a protein to be functionally relevant for several proteins, but irrelevant for a few others
Radiality[b]	Describes the integration of a node into the network	Highlights the ability of a protein to be functionally relevant for several proteins, but irrelevant for a few others
Centroid[b]	Describes the neighborhood of nodes by highlighting nodes that have the highest number of neighbors separated by the minimal shortest path	Highlights a protein that tends to be functionally capable of organizing discrete protein clusters or modules
Stress[b]	Describes the number of shortest paths that pass through a node	Highlights the relevance of a protein as functionally capable of holding together communicating nodes
Betweenness[b]	Describes, for each couple of nodes, the number of shortest paths that pass through a specific node	Highlights the relevance of a protein as functionally capable of holding together communicating nodes
Bridging[b]	Describes the neighborhood of nodes by highlighting nodes with a high number of high-degree neighbors	Highlights a protein possibly bringing in communication sets of regulatory protein
Eigenvector[b]	Describes a sort of weighted degree, where not only the number of the neighbors is important but also the Eigenvector of the neighbors itself	Highlights a protein interacting with several important proteins, suggesting a central super-regulatory role or a critical target of a regulatory pathways
Edge betweenness[c]	Describes, for each couple of nodes, the shortest paths that pass through a specific edge	Highlights the relevance of the interaction as capable of organizing regulatory process

For each centrality, it is described the topological and biological meaning. The [a] indicates network's properties. The [b] indicates node's properties. The [c] indicates edge's property

Another centrality measure widely used is known as betweenness [84]. It can be defined as follows: this measure summarizes the extent to which a vertex is located "between" other pairs of vertexes. In this case, centrality is based upon the perspective that importance relates to where a vertex is located with respect to the paths in the network graph. In other terms, betweenness centrality is based on communication flow. Nodes with a high betweenness centrality are interesting because they lie on communication paths and control information flow. Also called hubs/bottlenecks [85], they can represent important proteins in signaling pathways and can form targets for drug discovery. For example, by combining this data with interference analysis, targeted attacks on protein-protein interaction networks have been simulated to predict which proteins were better drug candidates [86].

Formally, betweenness can be defined as

$$\mathrm{Cl}(v) = \frac{\sigma(s,t|v)}{\sum_{s\neq t\neq v\in V}\sigma(s,t)} \tag{4}$$

where $\sigma(s,t|v)$ is the total number of shortest paths between s and t that pass through v, and $\sigma(s,t)$ is the total number of shortest paths between s and t (regardless of whether or not they pass through v).

Other centralities used to globally evaluate the structure of a network include:

- Degree distribution: a function describing the proportion of nodes related to each observed degree
- Modularity: evaluates the presence of modules, such as a group of nodes characterized by the tendency to form more connections within the group than outside [87]
- Cluster coefficient: the ratio of the number of edges among a node and its neighbors and the maximum possible number of edges among all of them [72]
- Motif/graphlet frequency: evaluates the presence of small subgraphs with a specific pattern that appear in a real-world network more frequently than in the relative random network [88]
- Edge clustering coefficient: the ratio between the number of triangles (three nodes connected by three edges) including an edge, and the maximum number of possible triangles may include the edge [89]
- Maximal Clique centrality: a property of a node taking into account the cliques (i.e, a subgraph in which each pair of nodes is connected) including the node [90]

The simplest way to perform a network topological analysis by evaluating these centralities is through Cytoscape's plugins, such as CentiScaPe [83] and NetworkAnalyzer [91], that provide the main basic methods to compute the topological properties of nodes, edges, and networks, both directed and undirected. Moreover, new plugins implementing recent developed topological centralities are CytoNCA [92] and CytoHubba [90].

4.2 Module analysis

Regardless of the approaches used to obtain a network, the detection of protein/gene modules is of great interest because they represent the functional units at the base of the mechanisms responsible of the cellular life. In biological networks, the term module has acquired three meanings: topological, functional, and pathological/disease. The analysis of the network structure allows to detect the topological modules defined as group of nodes highly interconnected [68]. These nodes are often related to well-defined molecular functions, thus, their detection PPI networks can help to identify functional modules [93], defined as a group of functionally related proteins/genes highly connected by genetic/physical interactions, co-expression, as well as membership of the same molecular complex or biological pathway [94]. The comparison between pathological and physiological conditions has finally led to the definition of disease modules, such as a set of nodes with a putative key role concerning mechanisms impaired due to disease [26, 51]. Topological, functional, and disease modules are generally not fully overlapped and often a single topological module can be linked to different functional or disease modules or vice-versa (Fig. 8).

Due to the complex connectivity of the biological networks, the identification of modules is a challenging task. Various methods have been proposed, and most of them are exclusively based on network topology. Some representative examples include the betweenness-based method [95], the modularity optimization method [96], the spectral partitioning method [97], the core-attachment based method [98], and the graph-theoretic approach relying on cliques [99] and other topological properties [100]. To improve the accuracy of module detection, the integration of functional information is more and more used [101–104]. These methods exploit the GO terms which in some cases are used to compute a similarity score that measures the edge weight and drives the module detection [105, 106].

The GO term enrichment analysis is routinely used also after the module detection to assess their biological relevance [107, 108]. Making use of statistical tests, these approaches evaluate if genes/proteins of a module are enriched in common functional properties (Fig. 9). During this process, standard methods treat each gene/protein as an isolated objects. However, in the last few years some network-based enrichment approaches have emerged taking into consideration also the interactions among molecules [109–111].

The commonly used methods for module detection have been extended to co-expression networks to evaluate

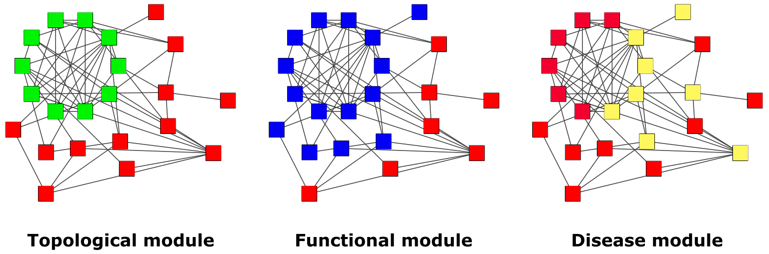


Fig. 8 Example of topological, functional and disease modules not fully overlapped. The *green* nodes indicate a topological module, the *blue* nodes indicate a functional module, while the *yellow* nodes indicate a disease module

the conditional patterns of co-expression and to provide insight into the cellular processes underlying the emergent phenotypes. Since genes could be co-regulated only across a subset of phenotypes, a biologically-motivated clustering method should be able to detect these patterns. This issue is faced by biclustering algorithms which clusterize both genes and experimental conditions. They are widely studied, and many different approaches have been published and applied to identify genes regulated in a state-specific manner [112].

In the context of module detection, the WGCNA package also provides a procedure consisting of a hierarchical clustering algorithm based on a distance matrix calculated by similarity measure between gene/protein pairs [59]. After assigning nodes to modules, an aggregate module signature, called eigenvector, is computed; it can be considered as an object representing the expression profiles of the molecules belonging to the module, thus, it simplifies the comparison of different modules [113]. A wide range of tools to perform module analysis are available. They include several Cytoscape's plugins, such as ClusterOne [114] and MCODE [100] and the Markov Cluster Algorithm (MCL) [115] or CFinder [99]. For a detailed view of these tools, the review by J.Ji et al. [116] is recommended.

5 Studies related to the use of protein co-expression networks

The investigation of proteomic data by co-expression-based approaches has been first addressed by Gibbs et al. to infer the protein abundance and to overcome issues linked to peptide-protein mapping [14]. Starting from experimental datasets obtained by LC-MS, and by using a method derived from WGCNA, the authors proposed a protein co-expression network approach (ProCoNa) where the nodes are peptides and the edges are calculated by processing their intensity. The modules computed by co-expression analysis were strictly correlated with the investigated phenotypes and showed a significant enrichment of some GO terms. Following these findings, the authors explored the relationship between co-expression networks reconstructed from transcriptomic and proteomic data [15]. In this study, concerning SARS-CoV infection, they used a bipartite graph analysis to evaluate

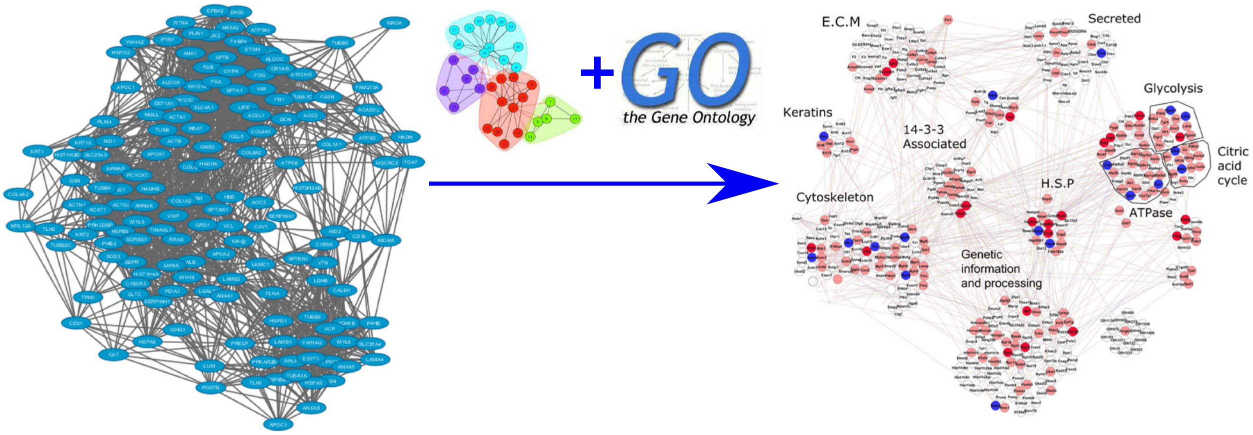


Fig. 9 Procedure used to identify/predict modules in biological networks. The network structure is used to identify groups of highly connected nodes by graph clustering algorithm, while the GO annotations are used to improve the accuracy of the cluster prediction. The final result are clusters of nodes highly connected and related to functions/processes significantly enriched, thus acting at the basis of the emergent phenotypes

phenotype associations, overlaps, and module correlation, thus, providing a foundation of a true multi-omics signatures.

The idea to use the WGCNA method on proteomic data was followed also by MacDonald et al. [18] to clarify the role of the glutamate signaling in schizophrenia (SCZ). The topological evaluation of the co-expression networks from SCZ affected subjects and healthy controls led to observe in SCZ affected group a lower average node degree. This result was probably due to the loss of coordination of the biological functions, as well as disease heterogeneity. However, in SCZ network, it was found the exclusive presence of a module enriched in GO terms related to glutamate signaling and whose proteins had a significant increased degree.

The application of the WGCNA on protein expression profiles was also faced by Chang Guo et al. to characterize the role of different protein isoforms in *E. Coli* resistance to serum killing [13]. Like in other cases, the authors evaluated the topological variations of the co-expression networks between control- and serum-treated groups. By considering the connectivity of modules identified in both networks, a protein, IleS, was found with a differential number of connections in control and treated groups. Of note, its involvement in the response to serum killing was confirmed by independent functional test based on a gene-deletion mutant, thus, confirming the utility to use protein co-expression networks also to identify putative drug targets.

To find phenotype-related biomarkers in the context of renal dysfunction, D. Wu et al. followed an approach based on the combination of differentially expressed proteins and PPI networks. For each pair of connected nodes they calculated the PC score, and the topological analysis of the reconstructed co-expression networks led to identify twelve proteins involved in the pathology [44]. Likewise, Yu et al. investigated the molecular mechanisms underlying the glioblastoma multiforme (GBM)[20]. They analyzed samples of macaque rhesus brain by both iTRAQ and RNA-seq approaches. The proteins identified were combined with STRING database and, for each experimentally validated PPI, the PC score was calculated using both protein and transcript levels. Since the PC score from proteomic data resulted significantly higher than score calculated using transcript levels, the authors focused on WGCNA to identify protein modules involved in the disease. Finally, a more detailed evaluation of these modules allowed the selection of eight genes of interest, and two of them were already known drug targets of GBM.

6 Conclusions

The aim of this review was to provide an overview on PPI and co-expression networks. In particular, presenting the recent idea of the protein co-expression networks and their use to infer biological knowledge by topological and module analysis. Although literature is yet too weak, protein co-expression networks represent a valid approach to obtain a novel overview of proteomic data and to provide new hypotheses about key molecules acting in pathophysiological states. Of course, its real value has to be assessed by further studies, but preliminary findings make it promising. The main limitation to perform the construction of protein co-expression networks may be attributed to the difficulty in measuring a proteome with enough coverage. A major consequence is the high rate of missing values that introduce loss of information and significant bias. In addition, batch effects may occur in datasets run in different days or by different technicians. Thus, data normalization is another key point in the context of proteomic data preprocessing. These aspects will be surely improved by future advances of the proteomic technologies which in recent years have received a big boost from genome sequencing and from the combination of liquid chromatography and mass spectrometry [117]. In any case, the availability of large-scale proteomic data already offers a new range of opportunities to improve the existing network models, and in particular PPI, in understanding the mechanisms behind the emergent phenotypes [8, 10, 108, 118, 119].

The results shown through the reviewed studies have evidenced a good relation between the topology of protein co-expression network and the emergent phenotypes. Like PPI networks, the characterization of hubs/bottlenecks and functional, topological and disease modules has proved to select the most important molecules. Despite these findings, statistical methods to construct co-expression networks by processing large-scale proteomic data still need to be improved [63, 64, 66]. To date, the available applications are mainly based on WGCNA framework, and studies evaluating other approaches are expected. Gaussian graphical models [120], partial correlation [121], or Bayesian networks [122] are more sophisticated approaches that are gaining favor over simple correlations due to their ability to separate direct from indirect variable associations. These methods need to use prior knowledge to estimate probabilistic interactions, and their implementation on typical -omics data may be computationally challenging due to the curse of dimensionality. However, they are widely adopted to integrate different -omics data [123, 124] and to infer transcriptional regulatory networks in the context of reverse-engineering processing techniques [48, 49].

Collection and integration of different -omics data represent essential points to perform a global evaluation of the biological systems and to improve the effectiveness of the current systems biology approaches. For these purposes, genomic and proteomic data are often used in combination with PPI networks. Since many studies are reporting a low direct correlation between mRNA and protein abundance [125, 126], their integration with molecules acting in the post-transcriptional regulation [127, 128] and metabolomic data [10] is necessary. In this scenario, PPIs and co-expression networks provide the possibility to apply a multi-omic strategy [15] that should improve level of significance in understanding biological mechanisms, including those related to diseases. Moreover, gene and protein co-expression networks give the opportunity to represent and to evaluate at system level including organisms that lack information about PPIs. In fact, except for human and other few organisms, PPIs are often inferred by homology making incomplete the theoretical models to describe the real-world networks, and with a connectivity affected by false positive interactions [129].

It is evident that the reconstruction of more complete and specific models is key to improve the current systems biology approaches. Molecules and interactions so far considered intracellularly should also be evaluated in tissues, and a new network of relationships that keeps in communication cells, tissues and organs will have to be considered too. On the other hand, computational tools are required to effectively integrate multi-omic experiments [130]. In addition to basic research, these improvements may have important effects into clinical applications opening the way toward the use of multiple biomarkers and their relationships [22–24]. They represent a chance to generate new mathematical models and algorithms for advanced diagnosis and prognosis methods which may lead toward a more preventive, predictive, and personalized medicine [27, 51]. These objectives are the major challenges to be addressed in the near future, and their achievement rely on the synergistic cooperation of biologists, physicists, mathematicians, and bioinformaticians.

Abbreviations

GBM: Glioblastoma multiforme; GEO: Geometric random graph; GO: Gene ontology; iTRAQ: Isobaric tag for relative and absolute quantitation; LC: Liquid chromatography; MS: Mass spectrometry; PC: Pearson's correlation; PPI: Protein-protein interaction network; ProCoNa: Protein co-expression network approach; SC: Spearman's correlation; SCZ: Schizophrenia; SRM: Selected reaction monitoring; TAP-MS: Tandem affinity purification coupled with mass-spectrometry; WGCNA: Weighted gene cp-expression network analysis; Y2H: Yeast two-hybrid

Acknowledgements

The study was supported by the Italian CNR FLAGSHIP InterOmics 2015 project funded by MIUR and by Italian Ministry of Economy and Finance for the Project 'FaReBio di Qualita'.

Authors' contributions

DV and DDS conceived the manuscript. DV, IZ, and DDS wrote the manuscript. DDS, PLM, and GM revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1] Institute for Biomedical Technologies - National Research Council (ITB-CNR), 93 Fratelli Cervi, Segrate, Milan, Italy. [2] Department of Computer Science, Systems and Communication DiSCo, University of Milano-Bicocca, 336 Viale Sarca, Milan, Italy.

Received: 18 November 2016 Accepted: 9 March 2017
Published online: 20 March 2017

References

1. D Petrey, B Honig, Structural bioinformatics of the interactome. Annu. Rev. Biophys. **43**, 193–210 (2014). doi:10.1146/annurev-biophys-051013-022726
2. H Kohestani, A Giuliani, Organization principles of biological networks: an explorative study. Biosystems. **141**, 31–39 (2016). doi:10.1016/j.biosystems.2016.01.004
3. Z Mousavian, J Díaz, A Masoudi-Nejad, Information theory in systems biology. part ii: protein-protein interaction and signaling networks. Semin. Cell Dev. Biol. **51**, 14–23 (2016). doi:10.1016/j.semcdb.2015.12.006
4. CE Mason, SG Porter, TM Smith, Characterizing multi-omic data in systems biology. J. Exper. Med. Biol. **799**, 15–38 (2014)
5. BFMSD Di Silvestre, P Mauri, in *Biomarker Validation, Technological, Clinical and Commercial Aspects*. Evaluation of Proteomic Data: From Profiling to Network Analysis by Way of Biomarker Discovery (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2015). doi:10.1002/9783527680658.ch9
6. V Mehta, L Trinkle-Mulcahy, Recent advances in large-scale protein interactome mapping. F1000Research. **29**(5) (2016). doi:10.12688/f1000research.7629.1
7. F Azuaje, Y Devaux, DR Wagner, Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network. BMC Syst. Biol. **4**, 60 (2010). doi:10.1186/1752-0509-4-60
8. RK Nibbe, M Koyutürk, MR Chance, An integrative -omics approach to identify functional sub-networks in human colorectal cancer. PLoS Comput. Biol. **6**(1), 1000639 (2010). doi:10.1371/journal.pcbi.1000639
9. J Nair, M Ghatge, VV Kakkar, J Shanker, Network analysis of inflammatory genes and their transcriptional regulators in coronary artery disease. PLoS ONE. **9**(4), 94328 (2014). doi:10.1371/journal.pone.0094328
10. C Procaccini, F Carbone, D Di Silvestre, F Brambilla, V De Rosa, M Galgani, D Faicchia, G Marone, D Tramontano, M Corona, C Alviggi, A Porcellini, A La Cava, P Mauri, G Matarese, The proteomic landscape of human ex vivo regulatory and conventional t cells reveals specific metabolic requirements. Immunity. **44**(2), 406–421 (2016). doi:10.1016/j.immuni.2016.01.028
11. Y Yu, S Li, H Wang, L Bi, Comprehensive network analysis of genes expressed in human oropharyngeal cancer. Am. J. Otolaryngol. **36**(2), 235–241 (2015). doi:10.1016/j.amjoto.2014.11.002
12. J Liu, L Jing, X Tu, Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. BMC Cardiovasc. Disord. **16**(1), 54 (2016). doi:10.1186/s12872-016-0217-3
13. Y Guo, Y Xing, Weighted gene co-expression network analysis of pneumocytes under exposure to a carcinogenic dose of chloroprene. Life Sci (2016). doi:10.1016/j.lfs.2016.02.074
14. DL Gibbs, A Baratt, RS Baric, Y Kawaoka, RD Smith, ES Orwoll, MG Katze, SK McWeeney, Protein co-expression network analysis (procona). J. Clin. Bioinforma. **3**(1), 11 (2013). doi:10.1186/2043-9113-3-11
15. DL Gibbs, L Gralinski, RS Baric, SK McWeeney, Multi-omic network signatures of disease. Front. Genet. **4**, 309 (2014). doi:10.3389/fgene.2013.00309
16. C Guo, X-J Liu, Z-X Cheng, Y-J Liu, H Li, X Peng, Characterization of protein species and weighted protein co-expression network regulation of escherichia coli in response to serum killing using a 2-de based proteomics approach. Mol. Biosyst. **10**(3), 475–484 (2014). doi:10.1039/c3mb70404a
17. D Wu, X Liu, C Liu, Z Liu, M Xu, R Rong, M Qian, L Chen, T Zhu, Network analysis reveals roles of inflammatory factors in different phenotypes of

kidney transplant patients. J. Theor. Biol. **362**, 62–68 (2014). doi:10.1016/j.jtbi.2014.03.006

18. ML MacDonald, Y Ding, J Newman, S Hemby, P Penzes, DA Lewis, NA Yates, RA Sweet, Altered glutamate protein co-expression network topology linked to spine loss in the auditory cortex of schizophrenia. Biol. Psychiatr. **77**(11), 959–968 (2015). doi:10.1016/j.biopsych.2014.09.006
19. EI Kanonidis, MM Roy, RF Deighton, T Le Bihan, Protein co-expression analysis as a strategy to complement a standard quantitative proteomics approach: Case of a glioblastoma multiforme study. PLoS ONE. **11**(8), 0161828 (2016). doi:10.1371/journal.pone.0161828
20. X Yu, L Feng, D Liu, L Zhang, B Wu, W Jiang, Z Han, S Cheng, Quantitative proteomics reveals the novel co-expression signatures in early brain development for prognosis of glioblastoma multiforme. Oncotarget (2016). doi:10.18632/oncotarget.7416
21. F Brambilla, F Lavatelli, D Di Silvestre, V Valentini, R Rossi, G Palladini, L Obici, L Verga, P Mauri, G Merlini, Reliable typing of systemic amyloidoses through proteomic analysis of subcutaneous adipose tissue. Blood. **119**, 1844–1847 (2012). doi:10.1182/blood-2011-07-365510
22. I Zoppis, M Borsani, E Gianazza, C Chinello, F Rocco, G Albo, AM Deelder, YEM Van Der Burgt, M Antoniotti, F Magni, G Mauri, in *BIOINFORMATICS 2012 - Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. Analysis of correlation structures in renal cell carcinoma patient data (SCITEPRESS, Setubal, 2012), pp. 251–256. doi:10.5220/0003855702510256
23. C Cava, I Zoppis, G Mauri, M Ripamonti, F Gallivanone, C Salvatore, M Gilardi, I Castiglioni, in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer (IEEE, US/Canada, 2013), pp. 608–611. doi:10.1109/EMBC.2013.6609573
24. C Cava, I Zoppis, M Gariboldi, I Castiglioni, G Mauri, M Antoniotti, Combined analysis of chromosomal instabilities and gene expression for colon cancer progression inference. J. Clin. Bioinforma. **4**(1) (2014). doi:10.1186/2043-9113-4-2
25. M Vidal, ME Cusick, A-L Barabási, Interactome networks and human disease. Cell. **144**(6), 986–998 (2011). doi:10.1016/j.cell.2011.02.016
26. M Gustafsson, CE Nestor, H Zhang, A-L Barabási, S Baranzini, S Brunak, KF Chung, HJ Federoff, A-C Gavin, RR Meehan, P Picotti, MÀ Pujana, N Rajewsky, KG Smith, PJ Sterk, P Villoslada, M Benson, Modules, networks and systems medicine for understanding disease and aiding diagnosis. Genome Med. **6**(10), 82 (2014). doi:10.1186/s13073-014-0082-6
27. E Guney, J Menche, M Vidal, A-L Barábasi, Network-based in silico drug efficacy screening. Nat. Commun. **7**, 10331 (2016). doi:10.1038/ncomms10331
28. O Mason, M Verwoerd, Graph theory and networks in biology. IET Syst. Biol. **1**(2), 89–119 (2007)
29. J De Las Rivas, C Fontanillo, Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol. **6**(6), 1000807 (2010). doi:10.1371/journal.pcbi.1000807
30. J Mintseris, Z Weng, Structure, function, and evolution of transient and obligate protein-protein interactions. Proc. Natl. Acad. Sci. U. S. A. **102**(31), 10930–10935 (2005). doi:10.1073/pnas.0502667102
31. ED Levy, CR Landry, SW Michnick, How perfect can protein interactomes be? Sci. Signal. **2**(60), 11 (2009). doi:10.1126/scisignal.260pe11
32. AG Ngounou Wetie, I Sokolowska, AG Woods, U Roy, JA Loo, CC Darie, Investigation of stable and transient protein-protein interactions: Past, present, and future. Proteomics. **13**(3-4), 538–557 (2013). doi:10.1002/pmic.201200328
33. D La, M Kong, W Hoffman, YI Choi, D Kihara, Predicting permanent and transient protein-protein interfaces. Proteins. **81**(5), 805–818 (2013). doi:10.1002/prot.24235
34. A Vinayagam, J Zirin, C Roesel, Y Hu, B Yilmazel, AA Samsonova, RA Neumüller, SE Mohr, N Perrimon, Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat. Methods. **11**(1), 94–99 (2014). doi:10.1038/nmeth.2733
35. BA Shoemaker, AR Panchenko, Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. PLoS Comput. Biol. **3**(4), 43 (2007). doi:10.1371/journal.pcbi.0030043
36. A Ceol, A Chatr Aryamontri, L Licata, D Peluso, L Briganti, L Perfetto, L Castagnoli, G Cesareni, Mint, the molecular interaction database: 2009 update. Nucleic Acids Res. **38**(Database issue), 532–539 (2010). doi:10.1093/nar/gkp983
37. S Kerrien, Y Alam-Faruque, B Aranda, I Bancarz, A Bridge, C Derow, E Dimmer, M Feuermann, A Friedrichsen, R Huntley, C Kohler, J Khadake, C Leroy, A Liban, C Lieftink, L Montecchi-Palazzi, S Orchard, J Risse, K Robbe, B Roechert, D Thorneycroft, Y Zhang, R Apweiler, H Hermjakob, Intact–open source resource for molecular interaction data. Nucleic Acids Res. **35**(Database issue), 561–565 (2007). doi:10.1093/nar/gkl958
38. A Franceschini, D Szklarczyk, S Frankild, M Kuhn, M Simonovic, A Roth, J Lin, P Minguez, P Bork, C von Mering, LJ Jensen, String v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. **41**(Database issue), 808–815 (2013). doi:10.1093/nar/gks1094
39. TS Keshava Prasad, R Goel, K Kandasamy, S Keerthikumar, S Kumar, S Mathivanan, D Telikicherla, R Raju, B Shafreen, A Venugopal, L Balakrishnan, A Marimuthu, S Banerjee, DS Somanathan, A Sebastian, S Rani, S Ray, CJ Harrys Kishore, S Kanth, M Ahmed, MK Kashyap, R Mohmood, YL Ramachandra, V Krishna, BA Rahiman, S Mohan, P Ranganathan, S Ramabadran, R Chaerkady, A Pandey, Human protein reference database–2009 update. Nucleic Acids Res. **37**(Database issue), 767–772 (2009). doi:10.1093/nar/gkn892
40. GD Bader, MP Cary, C Sander, Pathguide: a pathway resource list. Nucleic Acids Res. **34**(Database issue), 504–506 (2006). doi:10.1093/nar/gkj126
41. R Saito, ME Smoot, K Ono, J Ruscheinski, P-L Wang, S Lotia, AR Pico, GD Bader, T Ideker, A travel guide to cytoscape plugins. Nat. Methods. **9**, 1069–1076 (2012). doi:10.1038/nmeth.2212
42. Z Hu, J Mellor, J Wu, C DeLisi, Visant: an online visualization and analysis tool for biological interaction data. BMC Bioinforma. **5**, 17 (2004). doi:10.1186/1471-2105-5-17
43. Y Ding, M Chen, Z Liu, D Ding, Y Ye, M Zhang, R Kelly, L Guo, Z Su, SC Harris, F Qian, W Ge, H Fang, X Xu, W Tong, atbionet–an integrated network analysis tool for genomics and biomarker discovery. BMC Genomics. **13**, 325 (2012). doi:10.1186/1471-2164-13-325
44. J Wu, T Vallenius, K Ovaska, J Westermarck, TP Mäkelä, S Hautaniemi, Integrated network analysis platform for protein-protein interactions. Nat. Methods. **6**(1), 75–77 (2009). doi:10.1038/nmeth.1282
45. QIAGEN's Ingenuity pathway analysis. https://www.ingenuity.com/
46. G Wu, E Dawson, A Duong, R Haw, L Stein, Reactomefiviz: a cytoscape app for pathway and network-based data analysis. F1000Res. **3**, 146 (2014). doi:10.12688/f1000research.4431.2
47. D-Y Cho, Y-A Kim, TM Przytycka, Chapter 5: Network biology approach to complex diseases. PLoS Comput. Biol. **8**(12), 1002820 (2012). doi:10.1371/journal.pcbi.1002820
48. L Song, P Langfelder, S Horvath, Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinforma. **13**, 328 (2012). doi:10.1186/1471-2105-13-328
49. Z-P Liu, Reverse engineering of genome-wide gene regulatory networks from gene expression data. Curr. Genomics. **16**, 3–22 (2015). doi:10.2174/1389202915666141110210634
50. J Ruan, AK Dean, W Zhang, A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst. Biol. **4**, 8 (2010). doi:10.1186/1752-0509-4-8
51. A-L Barabási, N Gulbahce, J Loscalzo, Network medicine: a network-based approach to human disease. Nat. Rev. Genet. **12**(1), 56–68 (2011). doi:10.1038/nrg2918
52. CJ Wolfe, IS Kohane, AJ Butte, Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinforma. **6**, 227 (2005). doi:10.1186/1471-2105-6-227
53. Y Zhang, BR Fonslow, B Shan, MC Baek, JR Yates 3rd, Protein analysis by shotgun/bottom-up proteomics. Chem. Rev. **10**(113), 2343–94 (2013). doi:10.1021/cr3003533
54. B Usadel, T Obayashi, M Mutwil, FM Giorgi, GW Bassel, M Tanimoto, A Chow, D Steinhauser, S Persson, NJ Provart, Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ. **32**(12), 1633–1651 (2009). doi:10.1111/j.1365-3040.2009.02040.x
55. F Luo, Y Yang, J Zhong, H Gao, L Khan, DK Thompson, J Zhou, Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinforma. **8**, 299 (2007). doi:10.1186/1471-2105-8-299
56. LL Elo, H Järvenpää, M Oresic, R Lahesmaa, T Aittokallio, Systematic construction of gene coexpression networks with applications to

human t helper cell differentiation process. Bioinformatics. **23**(16), 2096–2103 (2007). doi:10.1093/bioinformatics/btm309

57. A Gobbi, G Jurman, A null model for pearson coexpression networks. PLoS ONE. **10**(6), 0128115 (2015). doi:10.1371/journal.pone.0128115
58. ExpressionCorrelation. http://www.baderlab.org/Software/ExpressionCorrelation
59. P Langfelder, S Horvath, Wgcna: an r package for weighted correlation network analysis. BMC Bioinforma. **9**, 559 (2008). doi:10.1186/1471-2105-9-559
60. P Langfelder, S Horvath, Fast r functions for robust correlations and hierarchical clustering. J. Stat. Softw. **46**(11), i11 (2012)
61. JD Storey, R Tibshirani, Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. **100**(16), 9440–9445 (2003)
62. B Zhang, S Horvath, A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. **4**, 17 (2005). doi:10.2202/1544-6115.1128
63. C Lazar, L Gatto, M Ferro, C Bruley, T Burger, Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. J. Proteome Res. **15**, 1116–1125 (2016). doi:10.1021/acs.jproteome.5b00981
64. L Nie, G Wu, DE Culley, JCM Scholten, W Zhang, Integrative analysis of transcriptomic and proteomic data: challenges, solutions and applications. Crit. Rev. Biotechnol. **27**, 63–75 (2007). doi:10.1080/07388550701334212
65. L Zhang, Y-Z Liu, Y Zeng, W Zhu, Y-C Zhao, J-G Zhang, J-Q Zhu, H He, H Shen, Q Tian, *et al*, Network-based proteomic analysis for postmenopausal osteoporosis in caucasian females. Proteomics. **16**(1), 12–28 (2016)
66. PC Carvalho, J Hewel, VC Barbosa, JR Yates, Identifying differences in protein expression levels by spectral counting and feature selection. Genet. Mol. Res. GMR. **7**, 342–356 (2008)
67. SWH Wong, N Cercone, I Jurisica, Comparative network analysis via differential graphlet communities. Proteomics. **15**(2–3), 608–617 (2015). doi:10.1002/pmic.201400233
68. M Girvan, MEJ Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci. U. S. A. **99**(12), 7821–7826 (2002). doi:10.1073/pnas.122653799
69. P Erdõs, On the evolution of random graphs. Publ. Math. Inst. Hungar. Acad. Sci. **5**, 17–61 (1960)
70. AL Barabasi, R Albert, Emergence of scaling in random networks. Science. **286**(5439), 509–512 (1999)
71. JP Josep Diaz, MD Penrose, M SERNA, Convergence theorems for some layout measures on random lattice and random geometric graphs. Comb. Probab. Comput. **9**, 489–511 (2000)
72. DJ Watts, SH Strogatz, Collective dynamics of 'small-world' networks. Nature. **393**(6684), 440–442 (1998). doi:10.1038/30918
73. R Albert, H Jeong, A-L Barabasi, Error and attack tolerance of complex networks. Nature. **406**(6794), 378–382 (2000). doi:10.1038/35019019
74. H Jeong, SP Mason, AL Barabási, ZN Oltvai, Lethality and centrality in protein networks. Nature. **411**(6833), 41–42 (2001). doi:10.1038/35075138
75. J-DJ Han, D Dupuy, N Bertin, ME Cusick, M Vidal, Effect of sampling on topology predictions of protein-protein interaction networks. Nat. Biotechnol. **23**(7), 839–844 (2005). doi:10.1038/nbt1116
76. N Przulj, DG Corneil, I Jurisica, Modeling interactome: scale-free or geometric? Bioinformatics. **20**(18), 3508–3515 (2004). doi:10.1093/bioinformatics/bth436
77. N Przulj, Biological network comparison using graphlet degree distribution. Bioinformatics. **23**(2), 177–183 (2007). doi:10.1093/bioinformatics/btl301
78. V Janjić, N Pržulj, The topology of the growing human interactome data. J. Integr. Bioinform. **11**(2), 238 (2014). doi:10.2390/biecoll-jib-2014-238
79. B Al-Anzi, P Arpp, S Gerges, C Ormerod, N Olsman, K Zinn, Experimental and computational analysis of a large protein network that controls fat storage reveals the design principles of a signaling network. PLoS Comput. Biol. **11**(5), 1004264 (2015). doi:10.1371/journal.pcbi.1004264
80. J-DJ Han, N Bertin, T Hao, DS Goldberg, GF Berriz, LV Zhang, D Dupuy, AJM Walhout, ME Cusick, FP Roth, M Vidal, Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. **430**(6995), 88–93 (2004). doi:10.1038/nature02555
81. P Tsaparas, L Mariño-Ramírez, O Bodenreider, EV Koonin, IK Jordan, Global similarity and local divergence in human and mouse gene co-expression networks. BMC Evol. Biol. **6**, 70 (2006). doi:10.1186/1471-2148-6-70
82. SL Carter, CM Brechbühler, M Griffin, AT Bond, Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics. **20**(14), 2242–2250 (2004). doi:10.1093/bioinformatics/bth234
83. G Scardoni, M Petterlini, C Laudanna, Analyzing biological network parameters with centiscape. Bioinformatics. **25**(21), 2857–2859 (2009). doi:10.1093/bioinformatics/btp517
84. H Wang, JM Hernandez, P Van Mieghem, Betweenness centrality in a weighted network. Phys. Rev. E Stat. Nonlinear Soft Matter Phys. **77**, 046105 (2008). doi:10.1103/PhysRevE.77.046105
85. X He, J Zhang, Why do hubs tend to be essential in protein networks? PLoS Genet. **2**, 88 (2006). doi:10.1371/journal.pgen.0020088
86. H Yu, PM Kim, E Sprecher, V Trifonov, M Gerstein, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput. Biol. **3**, 59 (2007). doi:10.1371/journal.pcbi.0030059
87. MEJ Newman, Modularity and community structure in networks. Proc. Natl. Acad. Sci. U. S. A. **103**(23), 8577–8582 (2006). doi:10.1073/pnas.0601602103
88. R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, U Alon, Network motifs: simple building blocks of complex networks. Science. **298**(5594), 824–827 (2002). doi:10.1126/science.298.5594.824
89. J Wang, M Li, H Wang, Y Pan, Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans. Comput. Biol. Bioinforma. **9**(4), 1070–1080 (2012). doi:10.1109/TCBB.2011.147
90. C-H Chin, S-H Chen, H-H Wu, C-W Ho, M-T Ko, C-Y Lin, cytohubba: identifying hub objects and sub-networks from complex interactome. BMC Syst. Biol. **8**(Suppl 4), 11 (2014). doi:10.1186/1752-0509-8-S4-S11
91. NT Doncheva, Y Assenov, FS Domingues, M Albrecht, Topological analysis and interactive visualization of biological networks and protein structures. Nat. Protoc. **7**, 670–685 (2012). doi:10.1038/nprot.2012.004
92. Y Tang, M Li, J Wang, Y Pan, F-X Wu, Cytonca: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. Bio. Syst. **127**, 67–72 (2015). doi:10.1016/j.biosystems.2014.11.005
93. V Spirin, LA Mirny, Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. U. S. A. **100**(21), 12123–12128 (2003). doi:10.1073/pnas.2032324100
94. LH Hartwell, JJ Hopfield, S Leibler, AW Murray, From molecular to modular cell biology. Nature. **402**(6761 Suppl), 47–52 (1999). doi:10.1038/35011540
95. MEJ Newman, M Girvan, Finding and evaluating community structure in networks. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. **69**(2 Pt 2), 026113 (2004). doi:10.1103/PhysRevE.69.026113
96. MEJ Newman, Fast algorithm for detecting community structure in networks. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **69**(6 Pt 2), 066133 (2004). doi:10.1103/PhysRevE.69.066133
97. L Donetti, MA Muñoz, Detecting network communities: a new systematic and efficient algorithm. J. Stat. Mech, P10012 (2004). doi:10.1088/1742-5468/2004/10/P10012
98. M Wu, X Li, C-K Kwoh, S-K Ng, A core-attachment based method to detect protein complexes in ppi networks. BMC Bioinforma. **10**, 169 (2009). doi:10.1186/1471-2105-10-169
99. B Adamcsek, G Palla, IJ Farkas, I Deré, T Vicsek, Cfinder: locating cliques and overlapping modules in biological networks. Bioinformatics. **22**(8), 1021–1023 (2006). doi:10.1093/bioinformatics/btl039
100. GD Bader, CWV Hogue, An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinforma. **4**, 2 (2003)
101. AL Hu, KCC Chan, Utilizing both topological and attribute information for protein complex identification in ppi networks. IEEE/ACM Trans. Comput. Biol. Bioinform. **10**(3), 780–792 (2013). doi:10.1109/TCBB.2013.37
102. S Srihari, HW Leong, A survey of computational methods for protein complex prediction from protein interaction networks. J. Bioinform. Comput. Biol. **11**(2), 1230002 (2013). doi:10.1142/S021972001230002X
103. X-F Zhang, D-Q Dai, L Ou-Yang, H Yan, Detecting overlapping protein complexes based on a generative model with functional and topological properties. BMC Bioinforma. **15**, 186 (2014). doi:10.1186/1471-2105-15-186

104. L Hu, KCC Chan, A density-based clustering approach for identifying overlapping protein complexes with functional preferences. BMC Bioinforma. **16**, 174 (2015). doi:10.1186/s12859-015-0583-3
105. J Wang, D Xie, H Lin, Z Yang, Y Zhang, Filtering gene ontology semantic similarity for identifying protein complexes in large protein interaction networks. Proteome Sci. **10**(Suppl 1), 18 (2012). doi:10.1186/1477-5956-10-S1-S18
106. M Kouhsar, F Zare-Mirakabad, Y Jamali, Wcoach: Protein complex prediction in weighted ppi networks. Genes Genet. Syst. **90**(5), 317–324 (2015). doi:10.1266/ggs.15-00032
107. A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, A Paulovich, SL Pomeroy, TR Golub, ES Lander, JP Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. **102**(43), 15545–15550 (2005). doi:10.1073/pnas.0506580102
108. F Brambilla, F Lavatelli, D Di Silvestre, V Valentini, G Palladini, G Merlini, P Mauri, Shotgun protein profile of human adipose tissue and its changes in relation to systemic amyloidoses. J Proteome Res. **12**(12), 5642–5655 (2013). doi:10.1021/pr400583h
109. C Zhang, J Wang, K Hanspers, D Xu, L Chen, AR Pico, Noa: a cytoscape plugin for network ontology analysis. Bioinformatics. **29**(16), 2066–2067 (2013). doi:10.1093/bioinformatics/btt334
110. A Alexeyenko, W Lee, M Pernemalm, J Guegan, P Dessen, V Lazar, J Lehtiö, Y Pawitan, Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinforma. **13**, 226 (2012). doi:10.1186/1471-2105-13-226
111. P Di Lena, PL Martelli, P Fariselli, R Casadio, Net-ge: a novel network-based gene enrichment for detecting biological processes associated to mendelian diseases. BMC Genomics. **16**(Suppl 8), 6 (2015). doi:10.1186/1471-2164-16-S8-S6
112. DJ Reiss, NS Baliga, R Bonneau, Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinforma. **7**, 280 (2006). doi:10.1186/1471-2105-7-280
113. P Langfelder, S Horvath, Eigengene networks for studying the relationships between co-expression modules. BMC Syst. Biol. **1**, 54 (2007). doi:10.1186/1752-0509-1-54
114. T Nepusz, H Yu, A Paccanaro, Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods. **9**(5), 471–472 (2012). doi:10.1038/nmeth.1938
115. S van Dongen, Graph clustering by flow simulation (2000). PhD thesis, University of Utrecht
116. J Ji, A Zhang, C Liu, X Quan, Z Liu, Survey: Functional module detection from protein-protein interaction networks. IEEE Trans. Knowl. Data Eng. **26**(2), 261–277 (2016). doi:10.1109/TKDE.2012.225
117. CC Tsou, D Avtonomov, B Larsen, M Tucholska, H Choi, AC Gingras, AI Nesvizhskii, Dia-umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods. **12**(3), 258–64 (2015). doi:10.1038/nmeth.3255
118. M Gstaiger, R Aebersold, Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat. Rev. Genet. **10**(9), 617–627 (2009). doi:10.1038/nrg2633
119. P Mauri, AM Riccio, R Rossi, D Di Silvestre, L Benazzi, L De Ferrari, RW Dal Negro, ST Holgate, GW Canonica, Proteomics of bronchial biopsies: galectin-3 as a predictive biomarker of airway remodelling modulation in omalizumab-treated severe asthma patients. Immunol. Lett. **162**(1) (2014). doi:10.1016/j.imlet.2014.08.010
120. S Ma, Q Gong, HJ Bohnert, An arabidopsis gene network based on the graphical gaussian model. Genome Res. **17**, 1614–1625 (2007). doi:10.1101/gr.6911207
121. L Han, J Zhu, Using matrix of thresholding partial correlation coefficients to infer regulatory network. Bio. Syst. **91**, 158–165 (2008). doi:10.1016/j.biosystems.2007.08.008
122. D Pe'er, Bayesian network analysis of signaling networks: a primer. Science's STKE Signal Transduct. Knowl. Environ. **2005**, 4 (2005). doi:10.1126/stke.2812005pl4
123. AR Joyce, BØ Palsson, The model organism as a system: integrating 'omics' data sets. Nat. Rev. Mol. Cell. Biol. **7**(3), 198–210 (2006). doi:10.1038/nrm1857
124. R Van Assche, V Broeckx, K Boonen, E Maes, W De Haes, L Schoofs, L Temmerman, Integrating -omics: Systems biology as explored through c. elegans research. J. Mol. Biol. **427**(21), 3441–3451 (2015). doi:10.1016/j.jmb.2015.03.015
125. G-W Li, XS Xie, Central dogma at the single-molecule level in living cells. Nature. **475**(7356), 308–315 (2011). doi:10.1038/nature10315
126. T Maier, M Güell, L Serrano, Correlation of mrna and protein in complex biological samples. FEBS Lett. **583**(24), 3966–3973 (2009). doi:10.1016/j.febslet.2009.10.036
127. R de Sousa Abreu, LO Penalva, EM Marcotte, C Vogel, Global signatures of protein and mrna expression levels. Mol. Biosyst. **5**(12), 1512–1526 (2009). doi:10.1039/b908315d
128. B Schwanhäusser, D Busse, N Li, G Dittmar, J Schuchhardt, J Wolf, W Chen, M Selbach, Global quantification of mammalian gene expression control. Nature. **473**(7347), 337–342 (2011). doi:10.1038/nature10098
129. X Peng, J Wang, W Peng, FX Wu, Y Pan, Protein-protein interactions: detection, reliability assessment and applications. Brief. Bioinform (2016). doi:10.1093/bib/bbw066
130. K Wanichthanarak, JF Fahrmann, D Grapov, Genomic, proteomic, and metabolomic data integration strategies. Biomark. Insights. **10**, 1–6 (2015). doi:10.4137/BMI.S29511
131. ELIXIR A distributed infrastructure for life-science information. http://160.80.34.9/elixir2015/