We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

4

Integrative Approach for Detection of Functional Modules from Protein-Protein Interaction Networks

Zelmina Lubovac-Pilav
University of Skövde, Systems Biology Research Centre
Sweden

1. Introduction

Advances in large scale technologies in proteomics, such as yeast two-hybrid (Y2H) screening and mass spectrometry (MS) have enabled us to generate large protein-protein interaction (PPI) networks. The structure of such networks has been frequently analysed to identify the modules, which constitute the basic "building blocks" of molecular networks. One of the challenges that systems biology is facing consists of explaining biological organisation in the light of the existence of modules in networks (Han et al., 2004; Pereira-Leal et al., 2004; Petti and Church, 2005; Rives and Galitski, 2003). A series of studies attempting to reveal the modules in cellular networks, ranging from metabolic (Ravasz et al., 2002), to protein networks (Spirin and Mirny, 2003; Yook et al., 2004), support the proposal that modular architecture is one of the principles underlying biological organisation.

Several key issues are being addressed in current research in systems biology, as a result of our post-genomic view that has expanded the role of the protein into an element of a network in which it has contextual functions within functional modules (Eisenberg et al., 2000; Jeong et al., 2001). How do modules interact to achieve a certain functionality (Han et al., 2004; Rives and Galitski, 2003)? How can we evaluate the biological relevance of modules (Pereira-Leal et al., 2004; Poyatos and Hurst, 2004)? Answering those questions may contribute to better understanding of the relationships between structure, function and regulation of molecular networks, which is an important aim of systems biology (Qi and Ge, 2006; Stelling et al., 2002).

From the structural perspective, modules are often associated with highly connected clusters of proteins. Many efforts in this area have been directed towards analysing structural properties of the protein interaction graph, measured by clustering coefficient and shortest path distance for example, to derive modular formations. The main focus presented in this chapter is on defining similarity between protein interactions based on an integrated score that takes into consideration topology of PPI network along with the functional knowledge determined by semantic similarity. An important reason for considering knowledge represented in annotations a valuable complement to topological characteristics is

encompassed in the concept of functional modules themselves. A functional module consists of proteins that cooperate towards achieving a particular function or participate in similar processes. Hence, considering annotation that describes molecular functions and biological processes should enrich the protein-protein interactions. Functional information can be retrieved from Gene Ontology (GO), which is a structured vocabulary used to annotate proteins with information about their molecular function, participation in biological processes or localization in cellular components. A module-identifying algorithm proposed earlier (Lubovac et al., 2006), SWEMODE (Semantic WEights for MODule Elucidation), that relies on an integrated measure, called semantic cohesiveness, corresponds to one of the successful approaches that contributes to achieve the important aims of systems biology. This method will be the focus of attention in this chapter.

2. Background

Molecular biology is becoming a highly modular science where functional modules are considered to be a critical level of biological organization. The term "module", as understood in molecular biology, was originally defined as a discrete unit with a function that is separable from those of other modules (Hartwell et al., 1999). Furthermore, modularity refers to clusters of elements that work in a co-operative fashion to achieve some defined function. Protein complexes constitute one example type of module, since the proteins within a complex interact functionally and physically to form a robust unit, which in its turn carries out some biological function (Yook et al., 2004).

One of the key issues to be solved with help of bioinformatics is the deciphering of the complex architecture of biological networks.

2.1 Climbing life's complexity pyramid

Biological networks are often modular and compound, and involve connections between groups of genes and proteins as well as between individual elements. A simple complexity pyramid (see Fig. 1) suggested by Oltvai and Barabasi (2002), illustrates different levels of cellular organisation.

Living systems are organised at both logical and physical levels. The individual nucleotides are elementary building blocks of DNA and RNA molecules, which, in turn, are organised into higher level structures such as regulatory elements, and genes. DNA is physically organised into larger structures such as chromatin and chromosomes. Groups of genes, proteins, RNAs (the bottom level of the pyramid in Fig. 1) may be organised into pathways in metabolism, and motifs in genetic regulatory networks (see level 2). Regulatory motifs may in turn serve as building blocks of functional modules (level 3). There is a growing body of evidence that the modules are then organised in a hierarchical manner (Barabasi and Oltvai, 2004; Oltvai and Barabasi, 2002; Ravasz et al., 2002), defining the large-scale functional organisation of the cell (level 4 in Fig. 1).

The way these various structures interact with each other determines the machinery of a cell. Cells and the extracellular matrix, which surrounds and supports cells, build up the tissues that in turn are organised into organs, and so forth.

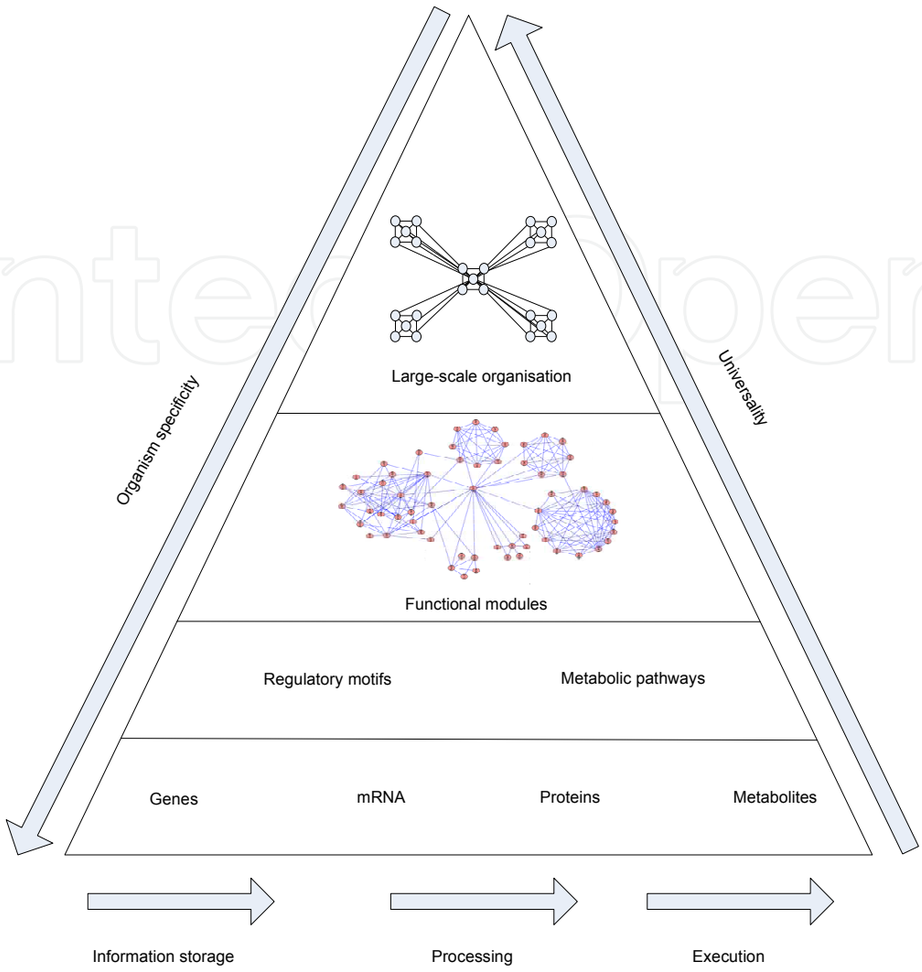


Fig. 1. Life's complexity pyramid redrawn from (Oltvai and Barabasi, 2002).

The integration of different layers in the pyramid to achieve a better understanding of system-level rules that govern cell function is one of the challenges in systems biology. Computational analysis tools and methods are needed at each level but also across different levels. Here, the integrative approach for deriving modules at the third level in the pyramid is described, which also make it possible to climb to the top, and provide means for revealing large-scale organisation.

2.2 Modularity in cellular networks

"Modularity is a fundamental design principle whereby components are partitioned according to common physical, regulatory, or functional properties" (Petti and Church, 2005). Modules can be found in many systems, for example, food webs, networks of web pages describing related subjects (Flake et al., 2002), networks of friends in sociology (Newman, 2003), or scientific collaboration networks (Newman, 2001). A usual synonym for the term module in other scientific disciplines, like sociology for example, is community or community structure. In a study by Flake et al., (2002), the term web community is for

example defined as "a collection of web pages such that each member page has more hyperlinks within the community than outside of the community". This definition may be adjusted further, according to Flake et al., (2002), to identify communities of varying sizes and levels of cohesiveness (clustering).

Furthermore, modularity involves groups of elements that work in a co-operative fashion to achieve some well-defined function. In a general network representation, a module appears as a highly interconnected group of nodes (Barabasi and Oltvai, 2004). Modules can be interpreted as separated substructures of a network or pathway, e.g. a protein complex is a module of a protein interaction network. Protein complexes are well-defined examples of modularity since they consist of proteins that interact functionally and physically to form a tightly connected unit, which, in turn, carries out some biological function (Yook et al., 2004). Another example of modular organisation can be found in genetic regulatory networks where several transcription factor binding sites, organised into functional units, i.e. modules, play a crucial role in gene transcription.

The members that constitute modules are more strongly related to each other than to members of other modules, which is reflected in the network topology. The modular nature of PPI networks is reflected by a high degree of clustering, measured by the clustering coefficient. The clustering coefficient measures the local cohesiveness around a node, and it is defined, for any node i, as the fraction of neighbours of i that are connected to each other (Watts and Strogatz, 1998). Simply stated, the clustering coefficient ci measures the presence of 'triangles' which have a corner at *i* (see the triangles with dashed sides in Fig. 2). The high degree of clustering is based on local sub-graphs with a high density of internal connections, while being less tightly connected to the rest of the network (Uhrig, 2006).

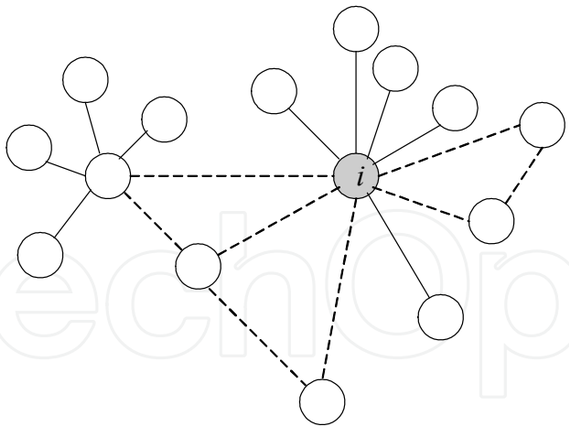


Fig. 2. Example of a protein sub-graph with triangle-forming proteins.

As pointed out by Barabasi and Oltvai (2004), each module may be reduced to a set of triangles, and a high density of such triangles is highly characteristic for PPI networks, pointing at the modular nature of such networks. By averaging the clustering coefficient over all nodes we can obtain a global measure of the cohesiveness of the network, where a high average clustering coefficient indicates the presence of modularity. It has been confirmed in many studies that most real large-scale networks tend to contain dense

clusters, in the sense that the average clustering coefficient of such networks is much greater than for random networks. In contrast, if modularity is absent in the network, the average clustering coefficient is comparable to that of a randomised network.

The exact meaning of modularity in biological networks depends on the network under consideration. For example, modules in protein networks are often seen as static molecular complexes (such as the ribosome) or as dynamic signalling pathways (such as the MAPK cascade). There are also examples of large modular molecule complexes that are in turn organised in modules. One of such complexes is yeast Mediator, which transmits regulatory signals from DNA-binding transcription factors to RNA polymerase II. The Mediator complex is thought to be composed of 24 subunits organised in four modules, named the head, middle, tail and Cdk8 modules. In gene regulatory networks, modules are often seen as sets of genes controlled by the same set of transcription factors under certain conditions (Segal et al., 2003).

Modules should not be seen as isolated components, since it has been shown that some crosstalk and overlap exists between them (Han et al., 2004; Schwikowski et al., 2000). Instead, modules should be considered as components that have dense intra-connectivity but sparse inter-connectivity. In a study analysing protein interaction networks in the yeast *Saccharomyces cerevisiae*, Schwikowski et al., (2000) reported global patterns of interactions of proteins within functional classes or subcellular compartments, as well as many possible cross-connections. It is further pointed out by Qi and Ge (2006) that the existence of the links between modules emphasises the coordination of the cellular processes. For example, Petti and Church (2005) investigated possible transcriptional coordination between glycolysis and lipid metabolism modules.

A growing body of work supports the idea that such modules underlie much of cellular functioning (Gavin et al., 2006; Han et al., 2004; Pereira-Leal et al., 2004; Qi and Ge, 2006; Rives and Galitski, 2003), and that functional modules are the most relevant organisational units of a cell from the perspective of systems biology (Hartwell et al., 1999).

2.3 Integrating functional knowledge in module discovery

Although topology-based network measures, such as clustering coefficient, play an important role in module discovery, there are some reasons why we should integrate functional knowledge as well when deriving modular formations. High-throughput protein interaction data that is often used to identify modules is very noisy (Titz et al., 2004). Technologies such as Y2H often result in many false positives that may cause false conclusions in the analysis. A possible approach to decrease the number of false interactions may be to focus on the "high confidence" data sets, where all interactions have been confirmed by several experiments. However, in this way the majority of the existing interactions would be discarded from further analysis. A better approach should imply incorporating the functional knowledge associated with available interactions into the analysis. This has also been pointed out in previous studies that focus on deriving protein complexes by using topological information. In (Przulj et al., 2004), it has been observed that the increasing size of PPI networks (by including medium and low confidence interactions) has resulted in a decreasing number of highly connected sub-graphs or clusters which may correspond to protein complexes. As Przulj, et al., (2004) state, the reason for this may be the

increasing noise in the data, and a possible solution to this problem is the integration of PPI networks with annotation or gene expression data. In sub-chapter 2.4 a possible general framework for such integrative approach for module identification is described.

2.4 A general framework for integrative module identification

There are many ways of measuring similarity between proteins. The main proposal presented here considers protein similarity based on an integrated score that takes into consideration protein interaction data (as a topology source) and functional information based on semantic similarity. As pointed out previously, an ideal approach should take into consideration both temporal and spatial data, to be able to reflect the true dynamics of the cellular networks. It is therefore worthwhile to discuss how the methods presented here may be generalised to cope with several sources of information. Our module-identifying framework may be generalised by:

1. considering several sources of topological information
2. considering several sources of functional information

Topological information may refer to, for example, protein-protein interactions obtained from different experimental sources, such as Y2H and MS. However, this information may also be derived from different topological properties like clustering coefficient, edge betweenness, etc.

Besides semantic similarity values based on protein GO terms that we used in this work, there are many other sources of functional information that may be useful for predicting membership in protein complexes. One of the most prominent sources is gene expression data generated using various high-throughput platforms, such as microarrays. Expression profile correlation coefficients may, for example, be used to assign similarity scores to pairwise interactions. Other sources of functional information are essentiality, phylogenetic profiles, localisation, the MIPS functional catalogue, etc.

In this study, as in the majority of others, protein interactions are treated as binary, i.e. the edges in a network are either present or absent. Bearing in mind the fact that large-scale methods, although offering vast improvements in efficiency, still have much higher error rates than small-scale methods, a step towards generalisation of the proposed algorithms would be to treat protein interaction networks probabilistically. By treating the edges as binary (indicating presence/absence of interaction), we cannot distinguish edges supported by multiple evidence types, from edges supported by evidence of differing quality. There are several ways of assigning probabilities to individual pairs of proteins based on the amount and type of supporting evidence (Asthana et al., 2004; Jansen et al., 2002; Jansen et al., 2003). When dealing with several data sources that need to be combined in order to improve the prediction, a usual way of combining these consists of overlapping different interactomes. This approach, in turn, gives rise to the question whether it is more beneficial to consider the union of the disparate datasets or their intersection. One of the extremes that may be envisaged is that each one of the networks that are to be integrated has a low rate of false positives (FP) but a high rate of false negatives (FN). In this case, the union of the two sets of interactions would be advantageous. At the other extreme, when dealing with networks with high FP rates and low FN rates, the intersection between the different networks is preferable.

The problem of finding an optimal combination of unions and intersections among the different networks may be defined, as described in (Jansen et al., 2002), as finding a trade-off between the highest possible coverage (TP/(TP+FN)) and the lowest possible error rate (FP/(TP+FP)). Determining the error rate is still an open question, as pointed out in (Jansen et al., 2002).

A hypothetical example of integrating different data sources that may be useful in generalising the proposed approaches is given in Fig. 3. The top part of the figure shows four possible data sources that may be useful for module identification. Two of them are topological sources, denoted as t_1 and t_2, and are usually treated as binary networks. The other two sources, denoted as f_1 and f_2, may be used to assign functional weights to the edges. For example, when using gene expression as a possible source for weighting the edges, the probability of finding two proteins in a complex, given a certain correlation between their expression profiles, may be a possible way to assign weights (Jansen et al., 2002). Gene ontology sub-graphs as a possible source of functional information is visualised in the third square in Fig. 3, where semantic similarity between ontology terms may be used to reflect the functional similarity between the proteins, as assumed in this work. These functional weights may also be transformed into binary values, by setting different thresholds, where the level of the threshold determines the sensitivity and specificity of the experiment.

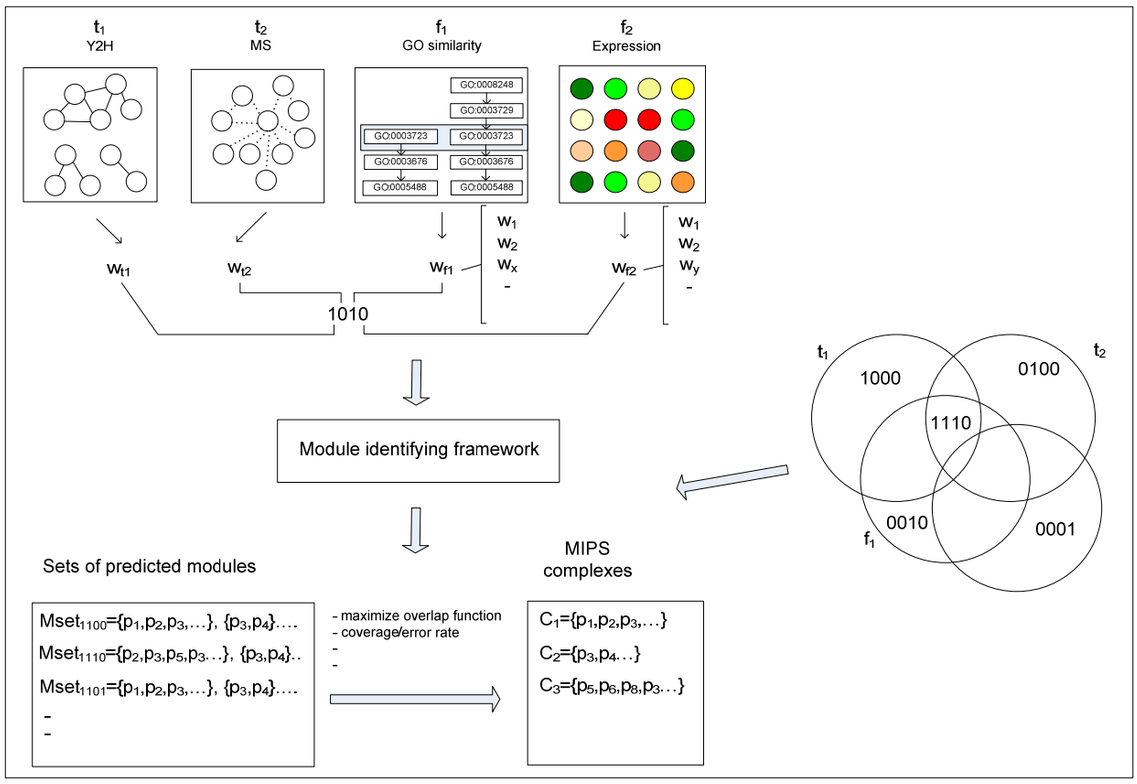


Fig. 3. Hypothetical integration of four data sources for module identification.

The bottom part of Fig. 3 shows the hypothetical module sets generated with different combinations of data sets. The Venn diagram to the right in the figure shows binary subset

profiles, where profile 1110 includes all data points that are present in data sets t_1, t_2, and f_1. Mset1110, for example, denotes the set of modules derived from the combination of MS, Y2H, and GO semantic similarity weights, where p_x denotes a protein x belonging the module.

3. Module identification based on an integrated approach

The algorithm described in previous work (Lubovac et al., 2006), SWEMODE (Semantic WEeights for MODule Elucidation), is an example of a method that employs an integrated approach for deriving functional modules, based on the functional and topological cohesiveness of the sub-graphs. Here, an integrated weighting score, called weighted clustering coefficient, that forms the bases for this method will be described. The reason for focusing on description of the integrative score here is that it can be applied as a part of node weighting procedure in other methods for deriving modules of PPI networks.

3.1 Weighted clustering coefficient

As depicted in earlier work, the separate edge weights do not provide an overall picture of the network's complexity. Therefore, we here consider the sum of all weights between a particular node and its neighbours, also referred to as the node strength. The strength s_i of the node i is defined as:

$$s_i = \sum_{\forall j, j \in N(i)} ss_{ij} \tag{1}$$

Given two proteins, i and j, with T_i and T_j containing m and n terms, respectively, the protein-protein semantic similarity ss_{ij} based on GO terms, is defined as the average inter-set similarity between terms from the given term sets (see Equation 2).

$$ss_{ij} = \frac{1}{m \times n} \sum_{t_k \in T_i, t_l \in T_j} sim(t_k, t_l) \tag{2}$$

Determining the similarity between two proteins i and j, is preceded by calculation of the similarity between the terms belonging to the term sets T_i and T_j that are used to annotate these proteins. Given the ontology terms $t_k \in T_i$ and $t_l \in T_j$, the semantic similarity measure proposed by (Lin, 1998) is defined as:

$$sim(t_k, t_l) = \frac{2 \ln p_{ms}(t_k, t_l)}{\ln p(t_k) + \ln p(t_l)} \tag{3}$$

Where $p(t_x)$ is the probability of term t_x and $p_{ms}(t_k,t_l)$ is the probability of the minimum subsumer of t_k and t_l, which is defined as the lowest probability found among the parent terms shared by t_k and t_l (Lord et al., 2003).

In previous work, some extensions of the topological clustering coefficient have been developed for weighted networks. In (Barrat et al., 2004), two scores that integrate topological and weighted features of the nodes – weighted clustering coefficient c^w and weighted average nearest-neighbours degree nn^w are introduced. These scores have

previously been applied to two types of complex weighted networks, namely, the world-wide airport network and the scientist collaboration network. A first attempt to apply these integrated scores on PPI networks was described in (Lubovac et al., 2006). A weighted measure that uses semantic similarity weights was introduced. Weighted clustering coefficient c^w is defined as:

$$c_i^w = \frac{1}{s_i(k_i - 1)} \sum_{\forall j,h \mid \{j,h\} \in K(i)} (ss_{ij} + ss_{ih}) \tag{4}$$

Where s_i is the functional strength of node i (see Equation 1) and ss_{ij} is the semantic similarity reflecting the functional weight of the interaction (see Equation 2). For each triangle formed in the neighbourhood of node i, involving nodes j and h, the semantic similarities ss_{ij} and ss_{ih} are calculated. Hence, not only the number of triangles in the neighbourhood of the node i is considered but also the relative functional similarity between the nodes that form those triangles, with regard to the total functional strength of the node. The normalisation factor $s_i(k_i\text{-}1)$ represents the summed weight of all edges connected from node i, multiplied by the maximum possible number of triangles in which each edge may participate. It also ensures that $0 \le c^w \le 1$. This measure can be involve any of the three aspects of Gene Ontology - molecular function, biological process and cellular component, or the combination of these.

4. Comparison with topology-based methods for module identification

The aim of this sub-chapter is to demonstrate the performance of the approach called SWEMODE (Lubovac et al., 2006), based on an integrative score described in 3.1, by comparing it to two purely topological approaches. One of the topology-based method for detecting modules from a PPI networks has been developed by Luo and Scheuerman (2006) and further analysed in (Luo et al., 2007). The module notion proposed was based on the degree definition of the sub-graphs. Unlike the approach described in Section 3, this method is based solely on topological properties of the protein sub-graph.

Modules generated with SWEMODE were also compared with the modules derived in (Przulj et al., 2004), based on HCS (Highly Connected Subgraphs) clustering algorithm (Hartuv and Shamir, 2000). This method aims to find disjoint subsets (clusters) that should satisfy following criteria: homogeneity – members of the same cluster are highly similar to each other; and separation: members of different clusters have low similarity to each other.

4.1 Protein-protein interaction data

For the evaluation purpose, two different PPI networks have been used. The first one was derived from the Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu), which is a database that stores and organises experimentally determined PPI (Xenarios et al., 2000). There is the subset of PPI from Yeast *S. cerevisiae*, denoted as CORE, which is the result of assessment with the Expression Profile Reliability Index (ERP Index) and the Paralogous Verification Method (PVM) (for further details, see (Deane et al., 2002)). The CORE subset contained 6379 interactions.

The second data set of PPI is obtained from the study by (von Mering et al., 2002). In that study, a quality assessment of large-scale data sets of protein-protein interactions in yeast was performed. A critical evaluation of the accuracy of high-throughput data is needed, because of the high rate of false interactions in these data sets. In (von Mering et al., 2002), data sets from yeast two-hybrid (Y2H) systems, protein complex purification techniques that rely on mass-spectroscopy (TAP and HMS-PCI), correlated mRNA expression profiles, genetic interactions, and *in silico* interaction predictions were analysed. As stated further in this study, each of these methods can be used to predict protein interactions, even though their goals are slightly different.

The authors integrated about 80 000 interactions between yeast proteins and found that only 2 455 were supported by more than one method. This low overlap between sets of protein interactions obtained from different methods may be due to the high fraction of false positives, but may also be caused by the difficulties for some methods to capture certain types of interactions. All interactions are classified by the level of confidence (low, medium, high), based on the evidence that supports them. In our study, we have used the interaction set with high level of confidence, meaning that all interactions are confirmed by several methods. This data set will be referred to as "von Mering". The data set contains 2 455 interactions between 988 proteins.

4.2 Evaluation against MIPS functional categories

The Munich Information Center for Protein Sequences (MIPS) provides high quality curated genome-related information, such as protein-protein interactions, protein complexes, protein functional categories, etc., spanning over several organisms.

The MIPS functional catalogue database consists of different fields, such as functional catalogue (FunCat) number, EC number, GO number, keywords etc. FunCat is an annotation scheme that provides functional descriptions of proteins (Ruepp et al., 2004). There are in total 28 main functional categories that are hierarchically structured. These categories cover functional fields such as metabolism, signal transduction, cellular transport etc.

The MIPS Comprehensive Yeast Genome Database (CYGD) provides information on the molecular structure and functional network of *S. cerevisiae*. The information used here for the evaluation purposes is the protein complex catalogue that contains a manually curated set of protein complexes that serve as an example of a type of module. There is another data set containing protein complexes obtained from (Gavin et al., 2002). This data set was produced by using a single experimental method, whereas the complex data set from MIPS has been derived from experiments from many labs using different techniques. Therefore, MIPS database is more realistic and appropriate to use for evaluation.

To evaluate and compare the performance of SWEMODE with two other methods for module identification, overlap score is used. In previous work, a similar evaluation has been applied to the clustering algorithm MCODE (Bader and Hogue, 2003), with respect to the number of matched complexes, but here slightly different definition of overlap score is used (see Equation 5).

The overlap score Ol (Poyatos and Hurst, 2004), is defined as:

$$Ol_{ij} = \left|M_i \cap M_j\right| \Big/ \sqrt{\left|M_i\right| \left|M_j\right|} \tag{5}$$

where M_i is the predicted module, and M_j is a module from the MIPS complex data set. The Ol measure assigns a score of 0 to modules that have no intersection with any known protein complex, whereas modules that exactly matches a known complex get the score 1.

4.3 Results

A total of 99 modules were detected in (Luo and Scheuermann, 2006). A new agglomerative algorithm was developed to identify modules from the network by combining the new module definition with the relative edge order generated by the Girvan-Newman algorithm. A JAVA program, MoNet, was developed to implement the algorithm Luo et al. (2007). Applying MoNet to the yeast core protein interaction network from the database of interacting proteins (DIP) identified 86 simple modules with sizes larger than 3 proteins. For convenience, those modules will be referred to as MoNet modules.

Evaluation of the MoNet modules with the overlap score threshold has been performed, and the results are compared with the resulting modules from SWEMODE, generated across approximately 400 different parameter settings (for parameter settings, see (Lubovac et al., 2006). We found that the modules derived from the latter show higher agreement with MIPS complexes (see Fig. 4). This comparison also indicates that introducing knowledge in terms

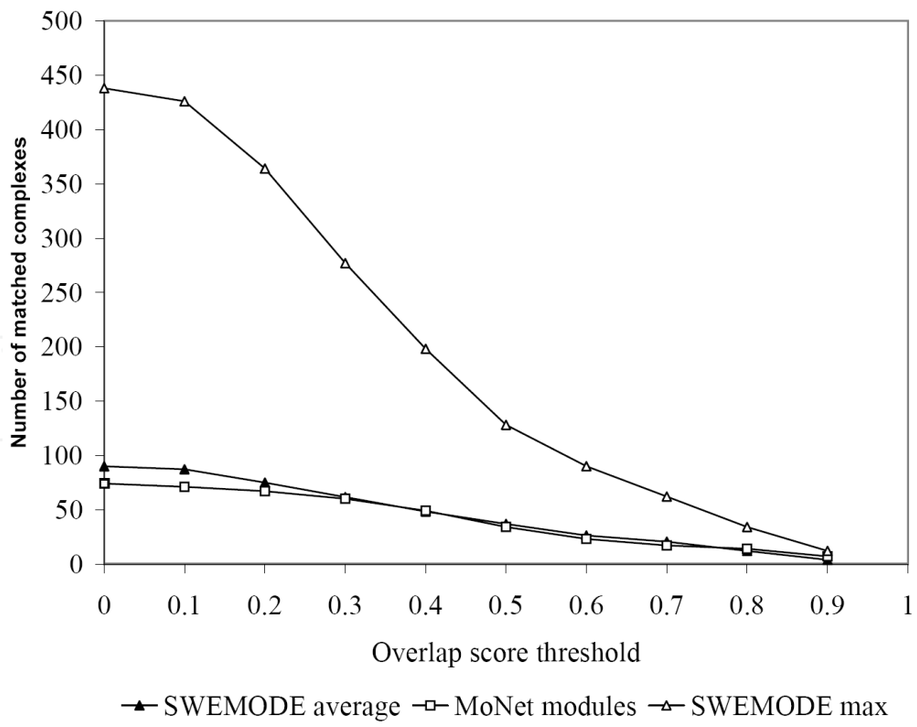


Fig. 4. Comparison between MoNet modules and SWEMODE modules.

of semantic similarity into the network topology seems to be advantageous over using only topology information. Furthermore, this method produces one single partition of the network, which does not seem biologically plausible, as many proteins may be involved in different processes.

We also compared our SWEMODE modules obtained from von Mering data with the modules derived in (Przulj et al., 2004), based on HCS. The modules generated with SWEMODE showed also here higher overlap with MIPS complexes (see Fig. 5). A more detailed analysis shows that both algorithms resulted in 39 identical modules. However, as HCS only discern the complexes that are highly interconnected, it discards many clusters that correspond to known complexes.

Another disadvantage of both methods that are here compared to SWEMODE is that they do not allow any overlap between modules, i.e. they produce disjoint clusters.

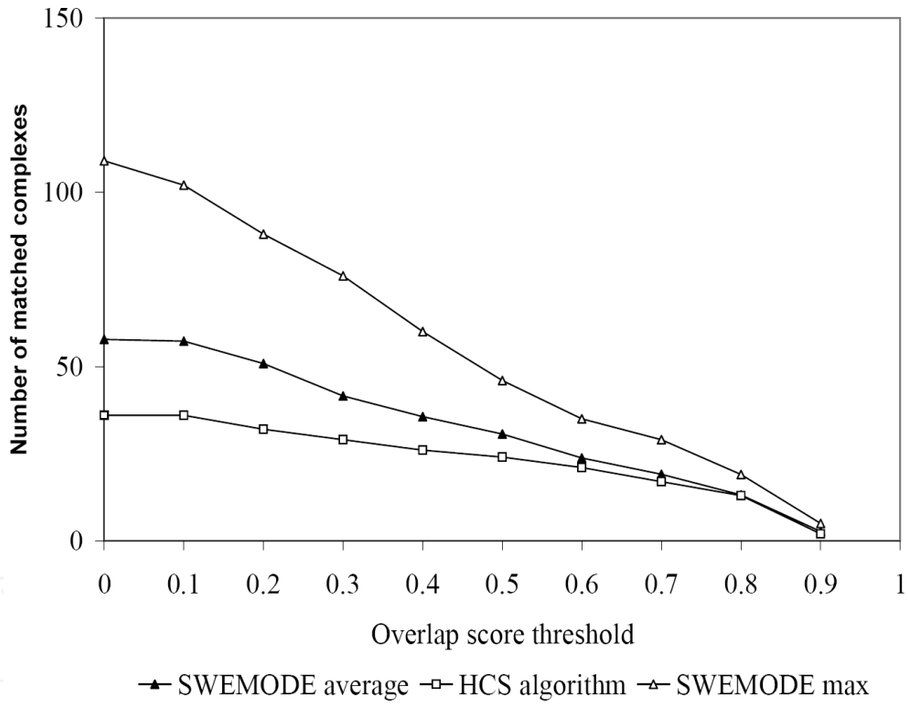


Fig. 5. Comparison between SWEMODE modules and modules generated with HCS clustering method.

5. Conclusion

The focus of attention in this chapter is the knowledge-based method that integrates domain specific knowledge, in this case functional information from Gene Ontology, with topological information, to derive modular structures from PPI networks. There are clear

disadvantages with the approaches that only rely on topological information, as previously described. In contrast to these methods that often suffer from lack of biological plausibility, the approach described here takes into consideration the functional knowledge about the experimental interactions, and in this way strengthen the validity of the obtained modular structures. Modules obtained in this way serve as models for studying interconnectivity, which is a step towards reconstruction of the higher order hierarchy of cellular networks.

Three different biological aspects – molecular function, biological process and cellular component, have been employed and tested for their suitability for deriving modules. The identification of protein complexes may become more challenging as additional PPI data becomes available, because the interactions are noisy, and the integration of PPI data with annotation might prove a useful solution to this problem. The integrated approaches contribute to this solution, by increasing the confidence in high-throughput Y2H data. The approach also provides means for an increased understanding of the higher-order structures underlying cellular function. As annotations become more complete, the increased biological relevance of our module predictions with integrated approaches is expected to be even more evident.

One of the biggest issues in this type of study is the difficulty to clearly characterise modules. There is no generally accepted definition of modules. A pioneering work in this area, performed by Hartwell et al. (1999) provides a wide definition, which leaves space for different authors to define different more specific criteria. This is, as also pointed out in (Schlosser and Wagner, 2004), unavoidable, and "retaining a pragmatic pluralism of different modularity concepts is probably a fruitful strategy for broadening our perspective and illuminating the importance of modularity at many different levels of organization".

A possible future application of the method described in this chapter is identification of modules of genes and proteins involved in various diseases, such as cancer. This module-level knowledge can contribute to the understanding of cancer on system-level, which may be useful for developing new drugs. Cancer-related networks for a specific type of cancer may be derived from, for example, gene expression data. Deriving gene networks makes it possible to apply network theoretic approaches on the interconnected genes that are potentially related to cancer development. Furthermore, a comparative analysis of the cancer-related networks derived from different types of cancer could be performed to identify modules that are shared among different types, but also to identify the specific processes that characterize a certain type of cancer.

Modular analysis may also be applied to identify general properties of the interrelated genes that are involved in the origin of cancer cells. A suitable model for this analysis is a gene fusion network in human neoplasia (Hoglund et al., 2006). By investigating topological properties of the cancer nodes in the network, such as node betweenness centrality, the cancer-related genes that act as "bridges" or communication points between various modules that correspond to cancer related processes may be identified.

Explaining the relationships between structure, function and regulation of molecular networks at different levels of the complexity pyramid of life is one of the main goals in systems biology. By integrating the topology, i.e. various structural properties of the

networks with the functional knowledge encoded in protein annotations, and also analysing the interconnectivity between modules at different levels of the hierarchy, we aim to contribute to this goal. With the increasing availability of protein interaction data and more fine-grained GO annotations, this will help constructing a more complete view of interconnected modules to better understand the organisation of cells.

6. References

Asthana, S., King, O. D., Gibbons, F. D., & Roth, F. P. (2004). Predicting protein complex membership using probabilistic network reliability. *Genome Res* 14, 1170-1175.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4, 2.

Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5, 101-113.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proc Natl Acad Sci U S A* 101, 3747-3752.

Deane, C. M., Salwinski, L., Xenarios, I., & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics* 1, 349-356.

Eisenberg, D., Marcotte, E. M., Xenarios, I., & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405, 823-826.

Flake, G. W., Lawrence, C., Giles, C. L., & Coetzee, F. M. (2002). Self-organization and identification of Web communities. *IEEE Computer* 35, 66-71.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* 440, 631-636.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P., & Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430, 88-93.

Hartuv, E., & Shamir, R. (2000). A clustering algorithm based on graph connectivity. *Information Processing Letters* 76, 175-181.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402, C47-52.

Hoglund, M., Frigyesi, A., & Mitelman, F. (2006). A gene fusion network in human neoplasia. *Oncogene* 25, 2674-2678.

Jansen, R., Lan, N., Qian, J., & Gerstein, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. *J Struct Funct Genomics* 2, 71-81.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* 302, 449-453.

Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.

Lin, D. (1998). An information-theoretic definition of similarity. *The 15th International Conference on Mashine Learning* (Madison, WI).

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275-1283.

Lubovac, Z., Gamalielsson, J., & Olsson, B. (2006). Combining functional and topological properties to identify core modules in protein interaction networks. *Proteins* 64, 948-959.

Luo, F., & Scheuermann, R. H. (2006). Detecting Functional Modules from Protein Interaction Networks. *Proceeding of the First International Multi-Symposiums on Computer and Computational Scineces* (IMSCCS'06) (IEEE Computer Society).

Luo, F., Yang, Y., Chen, C. F., Chang, R., Zhou, J., & Scheuermann, R. H. (2007). Modular organization of protein interaction networks. *Bioinformatics* 23, 207-214.

Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc Natl Acad Sci U S A* 98, 404-409.

Newman, M. E. J. (2003). Ego-centered networks and the ripple effect. *Socal networks* 25, 83-95.

Oltvai, Z. N., & Barabasi, A. L. (2002). Systems biology. Life's complexity pyramid. *Science* 298, 763-764.

Pereira-Leal, J. B., Enright, A. J., & Ouzounis, C. A. (2004). Detection of functional modules from protein interaction networks. *Proteins* 54, 49-57.

Petti, A. A., & Church, G. M. (2005). A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. *Genome Res* 15, 1298-1306.

Poyatos, J. F., & Hurst, L. D. (2004). How biologically relevant are interaction-based modules in protein networks? *Genome Biol* 5, R93.

Przulj, N., Wigle, D. A., & Jurisica, I. (2004). Functional topology in a network of protein interactions. *Bioinformatics* 20, 340-348.

Qi, Y., & Ge, H. (2006). Modularity and dynamics of cellular networks. *PLoS Comput Biol* 2, e174.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551-1555.

Rives, A. W., & Galitski, T. (2003). Modular organization of cellular networks. *Proc Natl Acad Sci U S A* 100, 1128-1133.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M., & Mewes, H. W. (2004). The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res* 32, 5539-5545.

Schlosser, G., & Wagner, G. P. (2004). Modularity in development and evolution: The University of Chicago Press).

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nat Biotechnol* 18, 1257-1261.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34, 166-176.

Spirin, V., & Mirny, L. A. (2003). Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 100, 12123-12128.

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., & Gilles, E. D. (2002). Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190-193.

Titz, B., Schlesner, M., & Uetz, P. (2004). What do we learn from high-throughput protein interaction data? *Expert Rev Proteomics* 1, 111-121.

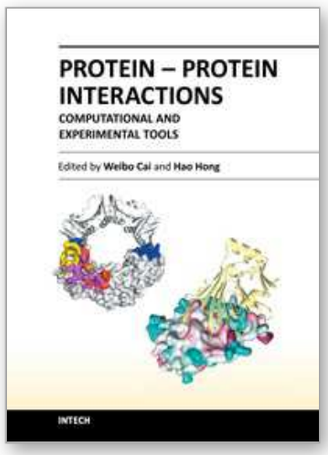
Uhrig, J. F. (2006). Protein interaction networks in plants. Planta 224, 771-781.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440-442.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399-403.

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., & Eisenberg, D. (2000). DIP: the database of interacting proteins. *Nucleic Acids Res* 28, 289-291.

Yook, S. H., Oltvai, Z. N., & Barabasi, A. L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* 4, 928-942.



Protein-Protein Interactions - Computational and Experimental Tools

Edited by Dr. Weibo Cai

ISBN 978-953-51-0397-4
Hard cover, 472 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

Proteins are indispensable players in virtually all biological events. The functions of proteins are coordinated through intricate regulatory networks of transient protein-protein interactions (PPIs). To predict and/or study PPIs, a wide variety of techniques have been developed over the last several decades. Many in vitro and in vivo assays have been implemented to explore the mechanism of these ubiquitous interactions. However, despite significant advances in these experimental approaches, many limitations exist such as false-positives/false-negatives, difficulty in obtaining crystal structures of proteins, challenges in the detection of transient PPI, among others. To overcome these limitations, many computational approaches have been developed which are becoming increasingly widely used to facilitate the investigation of PPIs. This book has gathered an ensemble of experts in the field, in 22 chapters, which have been broadly categorized into Computational Approaches, Experimental Approaches, and Others.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Zelmina Lubovac (2012). Integrative Approach for Detection of Functional Modules from Protein-Protein Interaction Networks, Protein-Protein Interactions - Computational and Experimental Tools, Dr. Weibo Cai (Ed.), ISBN: 978-953-51-0397-4, InTech, Available from: http://www.intechopen.com/books/protein-protein-interactions-computational-and-experimental-tools/integrative-approach-for-detection-of-functional-modules-from-protein-protein-interaction-networks

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen