Functional and Transcriptional Coherency of Modules in the Human Protein Interaction Network

Matthias E. Futschik[1], Gautam Chaurasia[1,2], Anna Tschaut[3], Jenny Russ[1],
M. Madan Babu[4] and Hanspeter Herzel[1]

[1] Institute for Theoretical Biology, Charité, Humboldt-University, Berlin, Germany
[2] Max Delbrück Center for Molecular Medicine, Berlin, Germany
[3] Department of Educational Science and Psychology, Freie Universität, Berlin, Germany
[4] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Summary

Modularity is a major design principle in interaction networks. Various studies have shown that protein interaction networks in prokaryotes and eukaryotes display a modular structure. A majority of the studies have been performed for the yeast interaction network, for which data have become abundant. The systematic examination of the human protein interaction network, however, is still in an early phase. To assess whether the human interaction network similarly displays a modular structure, we assembled a large protein network consisting of over 30,000 interactions. More than 670 modules were subsequently identified based on the detection of cliques. Inspection showed that these modules included numerous known protein complexes. The extracted modules were scrutinized for their coherency with respect to function, localization and expression, thereby allowing us to distinguish between stable and dynamic modules. Finally, the examination of the overlap between modules identified key proteins linking distinct molecular processes.

1 Introduction

Molecular interaction networks mirror the astonishingly complex interplay between numerous biological processes in living cells. To gain insights in these networks, major efforts have been undertaken to obtain comprehensive lists of interactions between biomolecules. Especially for the interactions between proteins, there has been a rapid growth of data due to large-scale screens, systematic review of literature and computational approaches. After initial efforts which targeted model organisms such as *S. cervisiae*, *D. melanogaster* and *C. elegans* (1-3), the assembly of the human interaction network has become a focal point of current research projects (4-10).

While the systematic mapping has produced a wealth of data, the elucidation of the underlying processes on a system-wide scale is still lagging behind. Major hurdles such as high false positive rates and experimental biases have to be overcome (11). Nevertheless, a few analyses in this direction have revealed interesting links between protein interactions and phenotypes for humans (12,13).

Many molecular functions require the interactions of several proteins. It is therefore important to identify *modules* of interacting proteins. Modules can be defined as clusters of proteins that are highly connected to each other, but sparely interact with the rest of the network. Modularity in protein interaction networks reflects both the tight interaction between proteins to perform a specific functions as well as the need for separation of interfering processes. Here, we aimed to gain an overview of the modular structures in the human protein interaction network. We merged several large literature-based interaction networks and identified subsequently tightly connected clusters of interacting proteins. Whereas previous studies concentrated on specific subsets of modules, our aim was the systematic assessment of coherency of function, localization and expression of the proteins in the identified modules.

2 Methods and Materials

2.1 Human protein-protein interaction data

Data on the human protein interaction network were collected from the Unified Human Interactome database (UniHI) (14,15). UniHI is a web-based publicly available database (www.mdc-berlin.de/unihi) which integrates human protein interaction data obtained from different sources. For our analysis, we extracted interactions included in the Human Protein Reference Database (HPRD), Biomolecular Interaction Network Database (BIND) and Database of Interacting Proteins (DIP) (4,8,16). These interactions were derived from the review of published literature. To ensure non-redundancy, we considered only interactions between proteins which could be mapped to their respective EntrezGene identifiers in the UniHI database. Altogether, over 35,000 interactions were extracted. Self- and redundant interactions were excluded from the obtained data leaving a total of over 31,000 interactions between more than 8,400 unique proteins for further analysis.

Note that we only considered binary interactions, i.e. direct interactions between proteins to ensure the reliability of the detected complexes. Complex interactions were excluded as they could otherwise interfere with the computational approach taken here for detection of modules.

2.2 Identification of modules in the protein interaction network

The identification of modules was based on the detection of *k-cliques*, i.e. fully connected subgraph of *k* vertices. Such *k*-cliques can form densely connected structures termed as *k*-clique *communitie*s. These communities are the union of all *k*-cliques that can be reached from each other through a series of adjacent *k*-cliques, where cliques sharing *k-1* nodes are defined as adjacent. Pella and co-authors previously developed a powerful tool Cfinder based on clique percolation method (CPM) for detecting overlapping *k*-cliques communities in networks (17). CPM first locates all *k*-cliques in a network and then identifies communities by carrying out standard component analysis of the clique-clique overlap. This method has been successfully applied to uncover the complex structure of overlapping communities in several types of networks (18). For our analysis, we applied Cfinder to identify highly connected modules in the human protein interaction network.

2.3 Generation of random graphs

To asses the significance of the identified cliques, we generated 100 random networks containing the same number of nodes and edges as in original network but with repeated random exchange of interactions. For instance, in such a procedure, two pairs of interacting proteins are randomly picked. The link between the nodes A and B (A-B) and between the nodes C and D (C-D) were changed to A-C and B-D, if such edges are not present in the original network. Note that since this is an undirected network, swapping of edges could happen between any pair of non-interacting nodes in the original network. Though there are several procedures to generate random networks, the current procedure which we adopted allows us to generate random networks with the same degree distribution as the original network. These random networks were used to obtain the expected number of cliques and were compared to the number of cliques obtained in the original interaction network.

2.4 Protein annotation

For the annotation of proteins, we utilized the Gene Ontology (GO) database supplying information about the assigned molecular function, biological process and cellular location (19). We assessed the significance whether the detected modules are enriched for proteins of certain functions, processes or locations by application of Fisher’s exact test. Since multiple testing was applied, the significance was adjusted by the Benjamini-Hochberg procedure delivering false discovery rates (20).

The coherency of modules with respect to cellular location was examined by an assessment of average pair-wise similarity of annotation of the participating proteins. To capture the similarity between two proteins, the induced GO graphs were compared. Subsequently, the size of their intersection divided by the size of their union was taken as a similarity measure (*simCC*). The values can range between 0 and 1 with larger values indicating larger similarity.

To facilitate the examination of localization of modules, we reduced the set of possible GO terms to so called informative categories. This previously introduced scheme selects GO categories which contain more than 100 genes while each of their children contains less than 100 genes (21) .

The GO analysis was carried out using the R/Bioconductor package *GO* and *GOstats* (22).

2.5 Expression data

To assess co-expression of proteins, we utilized a large human tissue expression dataset derived by 158 microarray measurements of 79 different tissue samples (23). Altogether, the expression level of over 16,000 genes was measured using Affymetrix HG-U133A and GNF1H arrays. Corresponding transcript levels were derived using Microarray Analysis Suite (MAS) and were provided by the authors. To improve the data consistency, we additionally applied quantile normalization. Using EntrezGene IDs, we could assign expression levels to approximately 8,000 proteins in our network. Co-expression was measured by the Spearman’s rank correlation.

3 Results

3.1 Identification of modular structures in the human interaction network

For the identification of modular structures, we applied the described CFinder algorithm for detection of *k*-cliques to the assembled human protein interaction network. Altogether, 671 distinct *k*-clique communities were detected with *k* ranging from 3 to 11 (see supplementary table S1). Most of the communities were based on 3- and 4-cliques (k = 3: 355; k = 4: 200). To assess the statistical significance, we constructed 100 random graphs with the same number of nodes and degree distribution and scrutinized them for the existence of cliques. Figure 1 shows the distribution of individual protein communities for different *k* in the original and random interaction networks. For k = 3, 4 and 5, similar numbers of cliques were found in random networks. However, for k = 6, only an average of 0.1 cliques were detected in the random networks, which is in sharp contrast to the 23 cliques found in the original network. Remarkably, no cliques of size larger than six were found in the random networks indicating the presence of a highly statistically significant modular structure in the human protein interaction network. This also confirms the findings in a previous study of the yeast

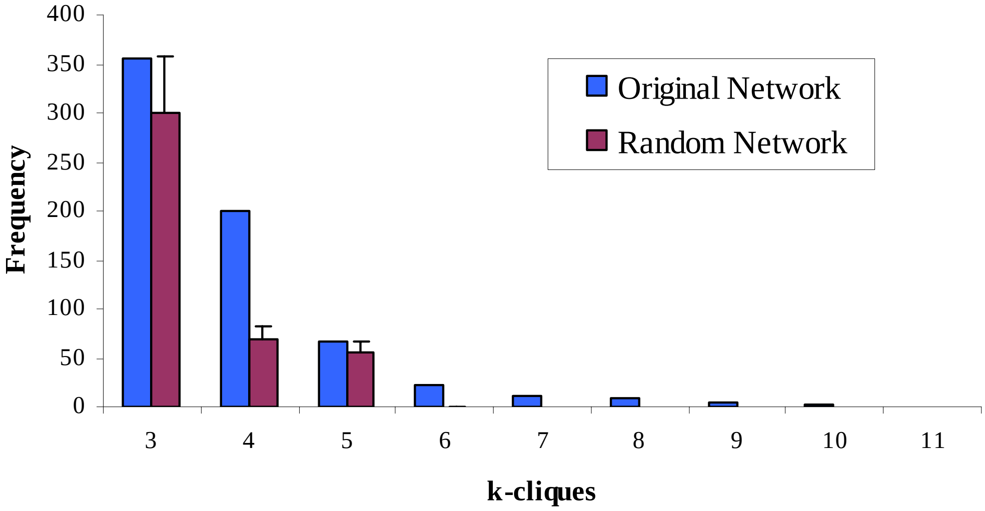


Figure 1: Identification of *k*-clique communities. The number of identified *k*-clique communities is shown for the original and random networks.

interaction network that highly interconnected enriched communities did not emerge by chance (24).

Next, we analysed the size of the individual communities for all *k*-cliques. As illustrated in supplementary figure S1, we found 267 *k*-cliques communities which have less than 5 proteins, most of them were based on 3- and 4-cliques. Only few communities included more than 15 proteins.

The number of communities in which a protein participates is highly variable (Supplementary figure S2). Most proteins are found in only one community (2,008) whereas TP53 - as a classical hub in protein interaction networks - is integrated in more than 20 different modules (Supplementary table S2).

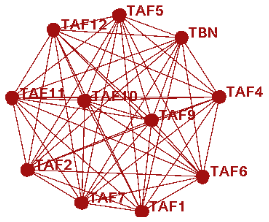
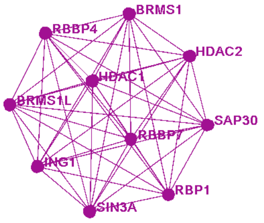
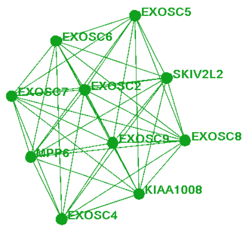
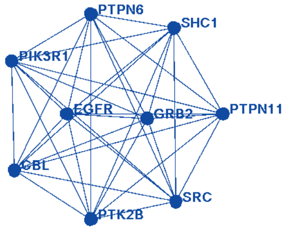
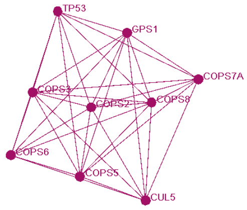
3.2 Functional annotation of the detected modular structures

We detected a large number of protein communities (or modules) based on *k*-cliques. But do these cluster structures reflect functional modules in the protein interaction network? To address this question, we used annotation information supplied by the Gene Ontology. Each detected modules was subsequently tested for enrichment of proteins assigned to specific GO categories. Examples of detected modules with annotation information are shown in table 1. To facilitate the interpretation, only GO categories are shown that were both significant and representative.

Many detected modules could be linked to known physical protein complexes. The largest identified module contained the TATA-binding protein (TBP) and multiple evolutionarily conserved TBP-associated factors (TAFs). The eleven included proteins are all known members of the transcription factor TFIID. Notably, this was also the largest fully connected clique discovered by Spirin and Mirny in the yeast interaction network (24).

Similarly, we can confidently link detected modules to the rRNA processing exosome complex and the COP9 signalsome, a highly conserved protein complex whose functions however are poorly understood. In contrast, modules were difficult to relate to known complexes if no prominent association with a specific cellular location existed.

Table 1: Examples of detected protein modules: *k* - size of cliques, *N* - number of proteins included in the module, *cor* - average correlation of expression. The false discovery rates are shown for representative biological processes and cellular components.

k	*N*	*Graph*	*Proteins*	*Biological process*	*Cellular component*	*Cor*
11	11	 TAF5 TAF12 TBN TAF11 TAF10 TAF4 TAF9 TAF2 TAF6 TAF7 TAF1	TAF1 TAF2 TAF10 TAF11 TAF12 TAF4 TAF5 TAF6 TAF7 TAF9 TBN	transcription initiation $6.14 \cdot 10^{-12}$	transcription factor TFIID complex $7.95 \cdot 10^{-26}$	0.39
10	10	 BRMS1 RBBP4 HDAC2 HDAC1 BRMS1L SAP30 RBBP7 ING1 RBP1 SIN3A	BRMS1 BRMS1L HDAC1 HDAC2 ING1 RBBP4 RBBP7 RBP1 SAP30 SIN3A	chromatin modification $5 \cdot 10^{-4}$	histone deacetylase complex $2.84 \cdot 10^{-05}$	0.26
9	10	 EXOSC5 EXOSC6 SKIV2L2 EXOSC7 EXOSC2 EXOSC9 EXOSC8 MPP6 KIAA1008 EXOSC4	EXOSC2 EXOSC4 EXOSC5 EXOSC6 EXOSC7 EXOSC8 EXOSC9 KIAA1008 MPP6 SKIV2L2	rRNA processing $3.10 \cdot 10^{-13}$	exosome $8.58 \cdot 10^{-19}$	0.30
9	9	 PTPN6 SHC1 PIK3R1 EGFR GRB2 PTPN11 CBL SRC PTK2B	CBL EGFR GRB2 PIK3R1 PTK2B PTPN11 PTPN6 SHC1 SRC	transmembrane receptor protein tyrosine kinase signalling $3.71 \cdot 10^{-9}$		0.11
7	9	 TP53 GPS1 COPS7A COPS3 COPS2 COPS8 COPS6 COPS5 CUL5	COPS2 COPS3 COPS5 COPS6 COPS7A COPS8 CUL5 GPS1 TP53		signalosome complex $6.86 \cdot 10^{-20}$	0.44

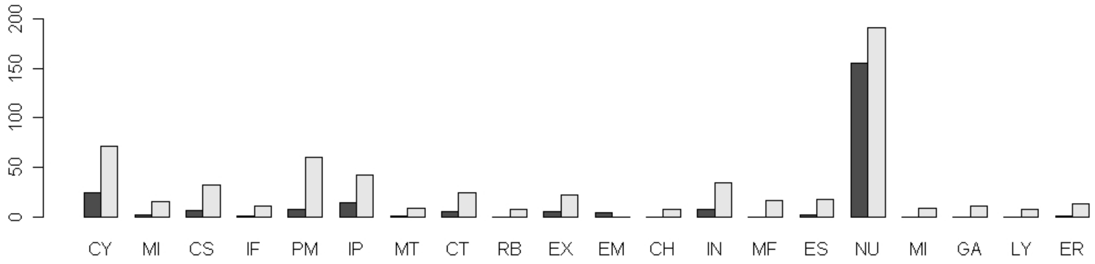


Figure 2: Sub-cellular localization of the detected modules. The number of modules is shown for which the majority (black bars) or a fraction of the included proteins (gray bars) was assigned to the corresponding cellular compartment. The distribution is based on the analysis of 316 modules which have a clique size *k* > 3. The following abbreviations are used: CY-cytoplasm, MI-mitochondrion, CS-cytoskeleton, IF-intermediate filament, PM-plasma membrane, IP-integral to plasma membrane, MT-microtubule, CT-cytosol, RB-ribosome, EX-extracellular region, EM-extracellular matrix, CH-chromosome, IN-intracellular, MF-membrane fraction, ES-extracellular space, NU-nucleus, MI-microsome, GA-Golgi apparatus, LY-lysosome and ER-endoplasmic reticulum.

3.3 Localization of modules

Previous analyses for yeast indicated that modules in interaction networks can be subdivided into protein complexes and dynamic functional modules (24). Protein complexes consist of tightly interconnected proteins which bind each other at the same time and location. In contrast, proteins in dynamic modules can interact at different times and locations despite being highly connected.

To analyse the co-location of proteins in the detected modules, we utilized information about their assigned cellular component in the GO. We reduced the set of possible GO terms to 20 informative categories to facilitate interpretation. Four categories comprised more than 1,000 proteins: ‘nucleus’ (3,895 proteins), ‘intracellular’ (1,931), ‘cytoplasm’ (1169) and ‘integral to plasma membrane’ (1,017).

Subsequent analysis showed a remarkably high degree of co-localization of proteins in modules. Of the 316 modules based on *k*-cliques (with *k* > 3), more than half (170) contained proteins allocated exclusively to a single cellular location. For over 75% of the modules, a majority of the included proteins were assigned to a single location.

Figure 2 displays the distribution of coherent locations of the modules. Most of the coherent modules were assigned to the nucleus (65%). Since proteins in steady complexes are necessarily co-localized, this observation may indicate an enrichment of protein complexes located in the nucleus.

3.4 Co-expression of modules

Besides the coherency of location, stable protein complexes might be distinguished from dynamic modules based on expression. We would expect that proteins in complexes underlie the same regulatory mechanism and thus would show co-expression. Of specific interest here is the question whether such co-expression correlates with the other distinct feature of complexes namely the co-localization of included proteins.

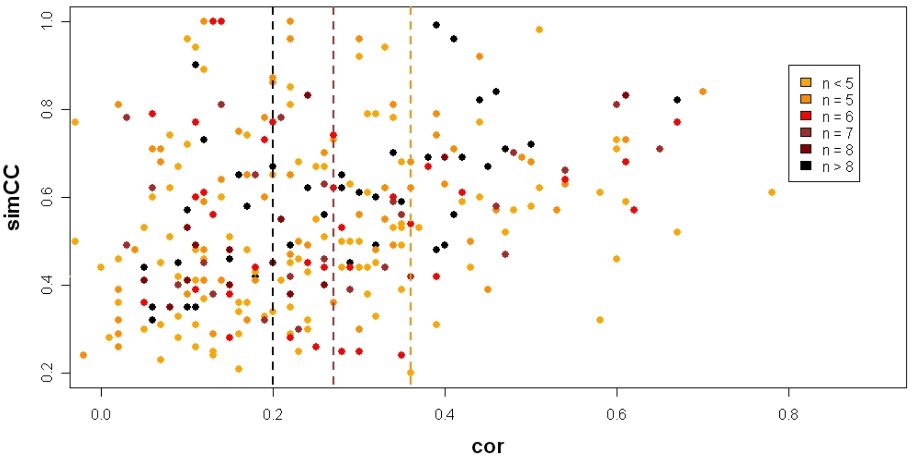


Figure 3: Coherency of co-expression and location within modules. The similarity of cellular localization (simCC) is plotted against the Spearman correlation. The size of the detected modules is colour-coded. The number *n* in the figure legend denotes the number of proteins included in the modules. Dashed lines indicate thresholds for different modules sizes where 99% of the correlation values in random samples are smaller for n = 5, n = 8 and n > 8.

To examine this issue, we calculated the correlation of expression within detected modules. The significance was assessed based on the expected correlation between randomly sampled proteins. Additionally, the similarity of cellular location based on GO annotation was derived (see section 2.4). Figure 3 displays both co-expression and similarity of location within modules. Comparison of the co-expression with co-localization of proteins within modules yields only a modest correlation of 0.27. This may indicate that a substantial percentage of the detected clusters in the interaction network are dynamic modules.

Inspection of this plot reveals that a majority of the modules containing 10 or more proteins is significantly co-expressed. In fact, 34 out of 51 modules (i.e. 66%) show a correlation coefficient larger than 0.20 for which 99% of equally sized random samples have smaller coefficients. Modules of smaller size are generally less significantly co-expressed due to a higher threshold for significance.

3.5 Overlap between modules and identification of linking proteins

Protein interaction networks are organized in multiple levels. Their lowest level is constituted by binding proteins to each other. These binding patterns can lead to the emergence of modular structures as we observed. Furthermore, the modules themselves can be interconnected by functional relationships. One major advantage of the applied algorithm for the detection of modules is that it allows modules to overlap. Thus, identified modules may constitute a higher level network. We exploited this possibility by creating a network of modules to analyse their functional relationship. Selecting modules based on 6-cliques, a highly connected network of 16 modules was detected (Figure 4). The largest module within this network contained over 80 proteins of which many are involved in signal transduction. Examples of the included proteins are members of the epidermal growth factor (EGF) receptor family (EGFR, ERBB2), janus kinases (JAK1, JAK2) and signal modifiers such as SOCS1. The second largest module of 51 proteins was enriched by various transcription factors such as the CREB-binding protein, forkhead box O1 (FOXO1), MYC, RB1 and TP53.

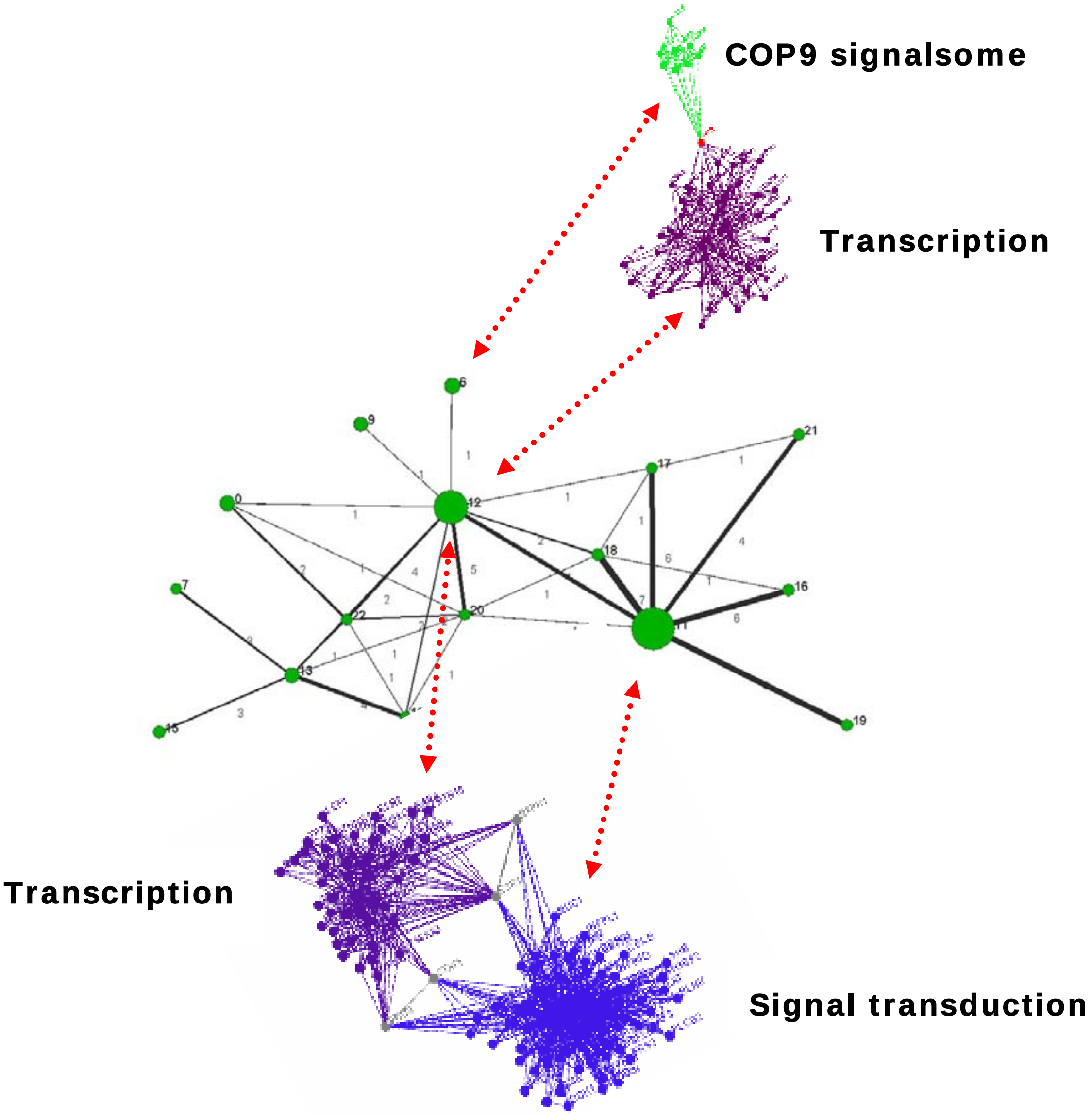


Figure 4: Network of modules. Nodes signify detected modules based on 6-cliques. The size of the nodes represents the number of proteins included in the corresponding modules. Edges between nodes indicate the existence of overlap. The width of the edges correlates with the number of linking proteins.

The association of the signal transduction module to the plasma membrane and the transcription module to the nucleus was highly significant (FDR = $6.00 \cdot 10^{-5}$ and $6.57 \cdot 10^{-21}$, respectively). Notably, these large modules are linked by four proteins (STAT1, STAT3, MAPK1, ESR1) which are known to shuttle between cytoplasm and nucleus.

In contrast, several modules were linked to the transcription module by single proteins. Examples of such sparse interconnections are the linkage of the transcription module to the COP9 signalosome complex by TP53 and to the TFIID complex by TBP.

4 Discussion

System-wide interaction network analysis offers the possibility to study cellular mechanisms in a comprehensive manner. However, there are numerous challenges to overcome. Interaction data are still sparse and might be compromised by a large number of false positives and by various experimental biases. In fact, we have recently demonstrated that the approach used for assembling protein interactions networks has severe effects on the resulting networks. For example, signalling proteins tend to be overrepresented in networks based on review of literature. Thus, it is not surprising that the largest module was associated with cell signalling since our network was constructed using only literature-based interactions maps. We utilized here only such interaction maps to facilitate the interpretation of the results. However, this restriction is likely to limit the number and type of possible modules that can be identified. Nevertheless, our study demonstrates clearly that the constructed human interaction network comprises a large number of functional modules. We further plan to incorporate the detected modules in the UniHI to facilitate the interpretation and usability of the human interactome.

Our analysis shows that many modules can be assigned to cellular processes. It also indicates that protein complexes and dynamic functional modules can be distinguished based on co-localization and co-expression, although there exists no rigorous threshold to distinguish them.

Note that the applied method for detection of modular structures is restrictive, since it requires fully connected cliques. Alternative methods may therefore be favourable to detect less densely connected modules. It should be noted that such restrictive definition of modules leads to an increased robustness of the detected modules regarding false positive interactions. Even if a substantial percentage of interactions are removed, the identified modules will still form highly connected clusters [24]. A further major advantage of the applied method is that an overlap between modules is allowed. This enabled us to identify potential key proteins linking different cellular processes. The constructed 'meta-network' of modules gives us a first intriguing image of the complex interplay between different components of the cellular machinery.

5 Acknowledgement

The presented work was supported by the SFB 618 grant of the *Deutsche Forschungsgemeinschaft (DFG).*

6 References

1. Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403,** 623-627.

2. Li, S. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303,** 540-543.

3. Giot, L. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302,** 1727-1736.

4. Bader, G.D. *et al.* (2001) BIND--The Biomolecular Interaction Network Database. *Nucleic Acids Res*, **29,** 242-245.

5. Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics*, **21,** 2076-2082.

6. Lehner, B. and Fraser, A.G. (2004) A first-draft human protein-interaction map. *Genome Biol*, **5,** R63.

7. Persico, M. *et al.* (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, **6 Suppl 4,** S21.

8. Peri, S. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, **13,** 2363-2371.

9. Rual, J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437,** 1173-1178.

10. Stelzl, U. *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122,** 957-968.

11. Futschik, M.E. *et al.* (2007) Comparison of human protein-protein interaction maps. *Bioinformatics*, **23,** 605-611.

12. Lage, K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, **25,** 309-316.

13. Goh, K.I. *et al.* (2007) The human disease network. *Proc Natl Acad Sci U S A*, **104,** 8685-8690.

14. Chaurasia, G. *et al.* (2007) UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, **35,** D590-594.

15. Chaurasia, G. *et al.* (2007) Flexible web-based integration of distributed large-scale interaction data sets. *Journal of Integrative Bioinformatics,* **4,** 51.

16. Xenarios, I. *et al.* (2000) DIP: the database of interacting proteins. *Nucleic Acids Res*, **28,** 289-291.

17. Adamcsek, B. *et al.* (2006) CFinder: locating cliques and overlapping modules in biological networks. *Bioinformatics*, **22,** 1021-1023.

18. Palla, G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435,** 814-818.

19. Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25,** 25-29.

20. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B*, **57,** 289-300.

21. Zhou, X. *et al.* (2002) Transitive functional annotation by shortest path analysis of gene expression data. *Proc Natl Acad Sci U S A*, **99,** 12783-12788.

22. Balasubramanian, R. *et al.* (2004) A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, **20,** 3353-3362.

23. Su, A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, **101,** 6062-6067.

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