GRAPH-THEORETICAL COMPARISON REVEALS STRUCTURAL DIVERGENCE OF HUMAN PROTEIN INTERACTION NETWORKS

MATTHIAS E. FUTSCHIK[1]
m.futschik@staff.hu-berlin.de

ANNA TSCHAUT[2]
tschaut@zedat.fu-berlin.de

GAUTAM CHAURASIA[1,3]
g.chaurasia@biologie.hu-berlin.de

HANSPETER HERZEL[1]
h.herzel@biologie.hu-berlin.de

[1]*Institute for Theoretical Biology, Charité, Humboldt-Universität, Berlin,Germany*

[2]*Department of Educational Science and Psychology, Freie Universität, Berlin, Germany*

[3] *Max-Delbrück-Center for Molecular Medicine, Berlin-Buch, Berlin, Germany*

Protein interactions constitute the backbone of the cellular machinery in living systems. Their biological importance has led to systematic assemblies of large-scale protein-protein interaction maps for various organisms. Recently, the focus of such interactome projects has shifted towards the elucidation of the human interaction network. Several strategies have been employed to gain comprehensive maps of protein interactions occurring in the human body. For their efficient analysis, graph theory has become a favourite tool. It can identify characteristic features of interaction networks which can give us important insights into the general structure of the underlying molecular networks. Although such graph-theoretical analyses have delivered us a variety of interesting results, their general validity remains to be demonstrated. We therefore examined whether independently assembled human interaction networks show common structural features. Remarkably, while some general graph-theoretical features were found, we detected a strong dependency of network structures on the method used to generate the network. Our study strongly indicates that graph-theoretical analysis can be severely compromised by the observed structural divergence and reassessment of earlier results might be warranted.

Keywords: Protein interaction; graph theory; human interactome

1. Introduction

As protein interactions are essential for cellular processes, their systematic identification has become an important target in molecular biology. Initial efforts to assemble comprehensive lists of interactions have been undertaken for model organisms such as *S. cerevisiae*, *D. melanogaster* and *C. elegans* (1-3). Recently, the elucidation of the human protein interactome (i.e. the complete set of protein interactions occurring in the human body) has become the major focus for many research groups (4-10). A variety of experimental and computational strategies to map the human interactome have been pursued. All of these approaches have their unique strengths and weaknesses. Currently, the major three strategies (with their advantages and disadvantages) are as follows (11):

- ***Literature-based interaction maps***. Protein interactions are derived from literature searches performed either by human experts or computational text-mining approaches. *Advantages*: i) This approach is not biased towards a specific experimental technique, ii) interactions are measured under a variety of conditions, and iii) maps include interactions that require post-transcriptional modifications specific to humans. *Disadvantages:* i) The false positive rates are difficult to estimate, and ii) the approach is highly biased towards proteins which are currently popular research targets.

- ***Orthology-based interaction maps.*** This approach is based on the assumption that protein interactions are evolutionarily conserved. Thus, interactions between proteins detected in other organisms are extrapolated to their human orthologs. *Advantages:* i) The method is entirely computational, enabling rapid and cost-effective construction of human interaction networks, and ii) it gains power through the abundance of interaction data for model organisms such as *S. cerevisiae*, *C. elegans*, *D. melanogaster* and *M. musculus*. *Disadvantages*: i) It is purely predictive, and ii) a considerable rate of false positives can arise through two types of errors: 1) the mapping to wrong orthologs, and 2) interactions are not conserved.

- ***Yeast two-hybrid (Y2H)-based maps.*** The Y2H method comprises on a screening approach using a set of modified proteins. Proteins are fused either to DNA-binding or transcriptional activator domains. Both types of fused proteins are subsequently co-expressed in yeast. If interaction occurs, a functional transcription factor (such as GAL4) is formed and a reporter gene is transcribed. *Advantages*: i) The Y2H assay enables systematic and rapid large-scale screening for interaction, ii) it is not biased towards known interactions, and iii) it can detect transient interactions. *Disadvantages*: i) Interactions are measured outside native surrounding (except for yeast proteins) and thus are unrelated to any physiological function, and ii) assayed proteins need to be located in the nucleus.

All of these mapping efforts have generated huge networks. For their analysis, graph theory has become an important tool. It has been applied to a variety of complex networks in various fields ranging from the World Wide Web to social networks. The aim of graph-theoretical analysis is the identification of characteristic network features and properties. In the field of systems biology, graph-theory has become a method of choice for the study of large interaction networks (12). Although the notation of molecular networks as simple graphs is clearly an oversimplification, the use of graph-theoretical tools has been demonstrated to be very useful for the general understanding of disease processes (13), internal network structures (14) and evolutionary processes (15).

Although the application of graph-theory to interaction networks has produced many intriguing findings, it is not clear if these results are of general validity or specific to the analysed network. We therefore critically examined whether independently assembled human interaction networks show common structural features.

2. Methods and Materials

2.1. *Construction of Protein-Protein Interaction Maps*

To assess whether current human interaction networks display common structures, we have selected eight interaction maps representing the three described approaches. These networks were subsequently scrutinized for common as well as differing structural features.

Two literature-based networks were assembled based on data from the Human Protein Reference Database (HPRD) and Biomolecular Interaction Network Database (BIND) (4,8). For construction of a third literature-based network (termed here as COCIT), we used a published list of interacting proteins derived by text-mining (16). Orthology-based networks were generated using data from the Online Predicted Human Interaction Database (OPHID) and HOMOMINT database (5,7). In addition, we selected an alternative collection of inferred interactions to build a third orthology-based interaction network (termed here as ORTHO) (6). Finally, two Y2H-based networks were derived from results of recently published Y2H screens for human protein interactions (9,10).

Note that all networks were independently assembled. For comparison, proteins were mapped to their corresponding EntrezGene IDs. The sizes of the generated networks are displayed in table 1. Further details can be found in references (11,17).

2.2. *Graph-theoretical measures*

For analysis, protein interaction maps were converted to graphs with proteins as nodes and interactions as links or edges. The resulting graphs can be characterized using a variety of graph-theoretical measures:

The most fundamental characteristic of a node in a graph is its number of links to other nodes. It is referred as the *degree* of a node. The *degree distribution P(k)* gives the fraction of proteins with *k* interactions in the total network. It can be used to distinguish different network classes. For example, the degree distribution follows a Poisson distribution for random networks of Erdös-Rényi type. Such networks have a typical node degree. In contrast, the power-law distribution ($P(k) \sim k^{-\gamma}$) is characteristic for the class of scale-free networks. The scale-free network architecture has been associated with robustness against failure of single components (12). A hallmark of scale-free topology is the appearance of so called network *hubs* i.e. highly connected nodes. The exponent γ determines the role of the hubs in the network. The smaller γ is, the larger the fraction of nodes connected to hubs is in the network.

The *shortest path* length between two nodes is defined as the minimum number of links included in the path between the nodes. For calculation, we used the shortest path algorithm by Dijkstra (18).The mean average path length of a network is the average

Table 1: Human protein interaction networks compared in this study. The following abbreviations were used: **P**- number of mapped proteins, **I** - number of mapped interactions, **AD** – average degree, **NN**- Number of sub-networks that include more than 1000, between 1000 and 101, between 100 and 10 or between 10 and 1 proteins, **MPL** - mean path length, **D** - network diameter, **γ** – degree exponent, **CC** - average clustering coefficient.

Network	P	I	AD	NN	MPL	D	γ	CC	Method
MDC-Y2H	1703	3186	3.7	1/0/38/4	4.9	13	1.63	0.01	Y2H-assay
CCSB-H1	1549	2754	3.5	1/0/90/27	4.4	12	1.46	0.05	Y2H-assay
HPRD	5908	15658	5.2	1/0/135/140	5	15	2.44	0.13	Literature
BIND	2677	4233	2.9	1/3/169/256	5.9	16	1.90	0.17	Literature
COCIT	3737	6580	3.5	1/7/545/0	5.9	20	2.18	0.43	Literature
OPHID	2284	8962	7.8	1/3/95/0	4.8	15	1.36	0.23	Orthology
ORTHO	3503	9641	5.4	1/2/183/9	6.5	17	2.14	0.19	Orthology
HOMOMINT	2556	5582	4.2	1/0/85/45	5.1	12	2.76	0.07	Orthology

shortest path lengths between all possible pairs of proteins. The network's diameter is the maximum shortest path between two nodes included. To measure the local tendency of neighbors to be linked, the clustering coefficient can be utilized (12). It is defined as $C=2n/m(m-1)$ where n is the number of links between m neighbors. A large clustering coefficient indicates that neighbors of a node are likely to interact to each other.

To avoid artifacts, self-interactions were excluded in the graph-theoretical analysis and all calculations were performed based on the largest connected graph for each map. The analysis was carried out in the R language using the Bioconductor packages *graph* and *GraphAT* (19,20).

2.3. *Generation of random graphs*

We assessed the significance of the results by comparison with two background network models: i) Random graphs with the same number of nodes and interactions, but without conservation of the degree distribution. Interactions were assigned to randomly selected pairs of proteins until the random graph had the same number of interactions as the original network. ii) Random graphs with conservation of number of nodes and interactions as well as of the degree distribution. To construct such graphs, we started with the original network and repeatedly exchanged interactions in a random manner: Edges between node A and B (A-B) and between C and D (C-D) will be changed to A-C and B-D, if such edges are not present yet. Thus, the degree of A, B, C and D is conserved, whereas the connections are changed.

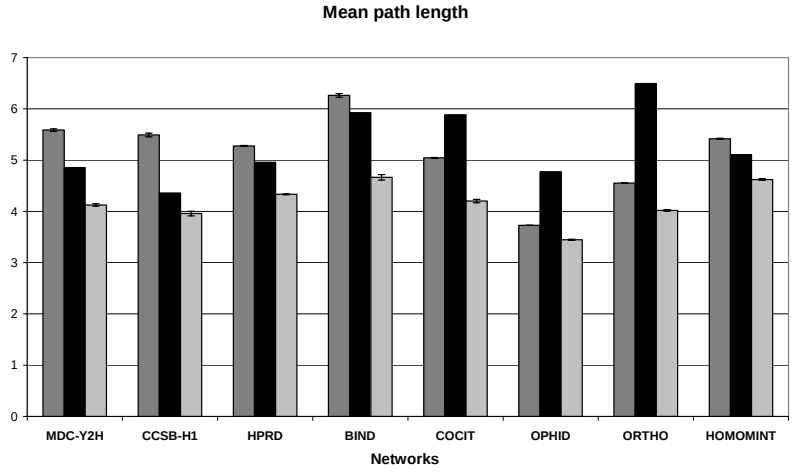

Figure 1: Mean path lengths of interaction networks: Black bars correspond to original graphs, dark gray bars correspond to random graphs with the same number of proteins and interactions and light gray bars correspond to random networks with conserved degree distribution. Errors bars show the standard deviations derived for three independent randomizations.

3. Results

3.1. *Connectivity of networks*

Using graph-theoretical measures, fundamental topological properties of protein interaction maps can be compared and characterized. After converting all interaction maps to graphs, we analyzed their internal connectivity (table 1). For all graphs, the vast majority of proteins were connected in a main network, which appears to be a general feature of protein-protein interaction networks, being also observed in other species (1-3,21). The remaining proteins formed predominantly smaller networks of less than 10 proteins. Only for BIND, COCIT, OPHID and ORTHO, medium sized networks (including 100-1000 proteins) emerged. Whether such separated 'islands' are artifacts reflecting the fragmentary state of proteins maps or functionally separated units remains subject for further research.

3.2. *Small worlds*

A main conclusion of previous studies was that protein interaction networks display 'small world' properties having small mean path lengths. This is also the case for the networks compared here (table 1 and fig. 1). Their mean path length is similar and ranges from 4.4 (CCSB-H1) to 6.5 (ORTHO). For most networks, it is smaller than expected for the corresponding random graphs. For all networks, however, the mean path length is

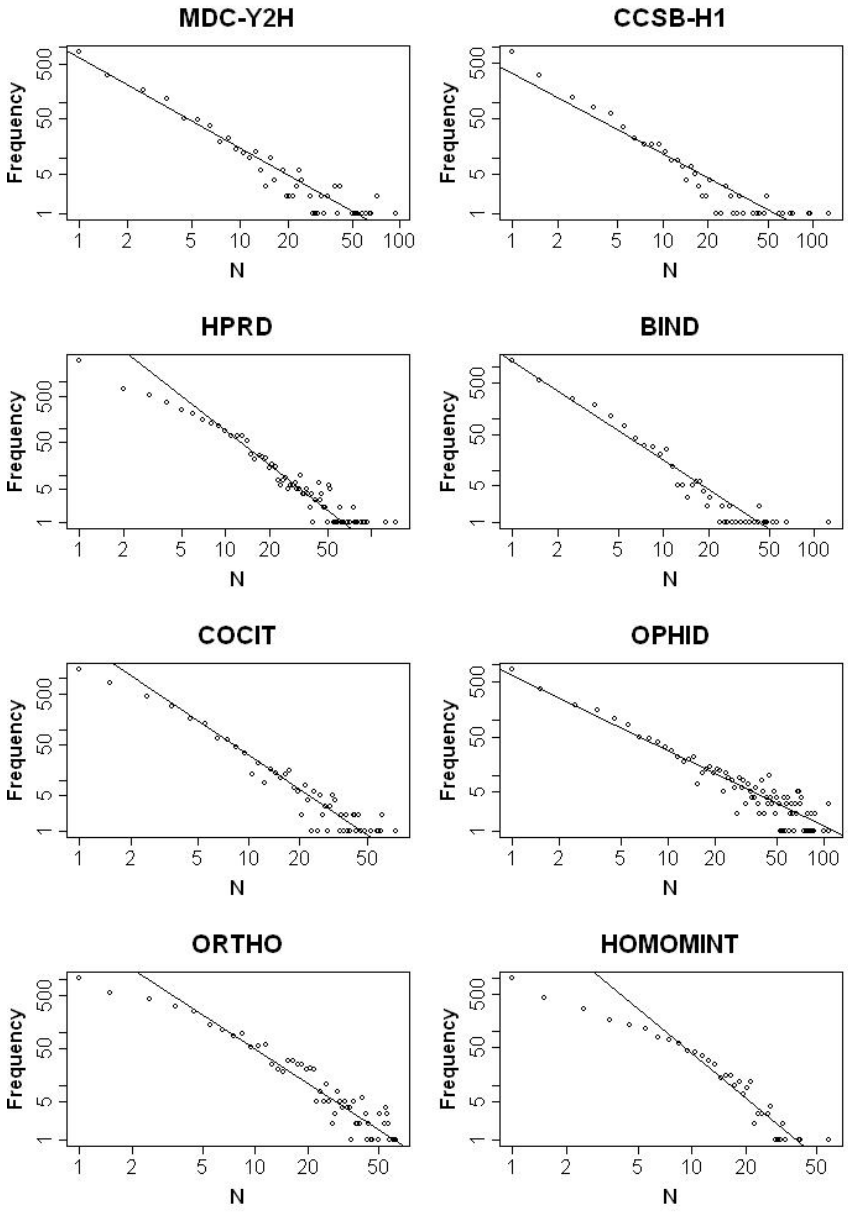

Figure 2: Degree distributions. The number of proteins was plotted as a function of the number of neighbors that proteins in the interaction maps have. For all maps, the degree frequencies follow a power-law $P(k) \sim k^{-\gamma}$ with some derivations for HPRD, COCIT, ORTHO and HOMOMINT. The exponent γ was derived by linear regression of the logged data.

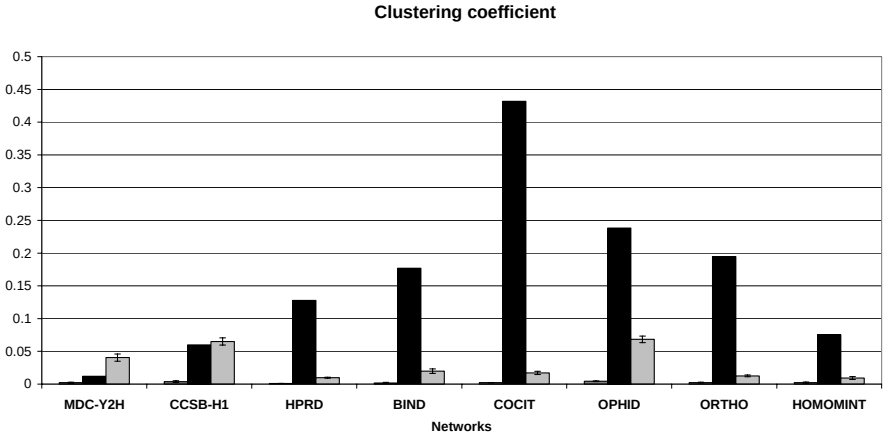

Figure 3: Mean clustering coefficient of interaction networks. The same representation as in figure 1 was used.

larger than expected for the corresponding random scale-free networks, pointing to the existence of internal structures (see also *Supplementary Materials*). The diameter (i.e. the largest path length within a network) ranges noticeably between 12 (CCSB-H1, HOMOMINT) and 20 (COCIT).

3.3. *Degree distribution*

An important determinant of a network's structure is the degree distribution *P(k)*. We found that all networks display power-law distribution implying a general emergence of hubs (fig. 2). However, some deviations can be observed. Networks derived from BIND, OPHID or Y2H-assays followed most closely the power-law distribution, in contrast the remaining ones show a relative depletion of interaction-poor proteins. Furthermore, the exponent γ varies by a factor of two between networks indicating that the role of hubs differs considerably (see section 2.2). Notably, networks that obey closely the power-law distribution also tend to have smaller mean path lengths.

3.4. *Modularity*

Cellular networks have been proposed to exhibit modular structure i.e they can be divided into separable highly connected sub-networks (12,14). A commonly used measure for modularity is the clustering coefficient reflecting the cohesiveness of the neighborhood of network nodes. In our analysis, the average clustering coefficient ranges remarkably by a factor of almost 50 from 0.01 to 0.45 (fig. 3). The smallest coefficients were found for Y2H-based networks; they were similar to the expected values for random scale-free networks, leading to the conclusion that the Y2H-based maps do not display particularly strong neighborhood cohesiveness. A possible reason could be a large number of undetected interactions (false negatives). In contrast, clustering

coefficients for literature- and orthology-based networks were considerably larger than for the corresponding random networks, implying that these networks are highly modular.

3.5. *Hierarchical structure*

Besides the assessment of modularity, the clustering coefficient has been employed to study hierarchical modular structures of networks. The concept of hierarchical modularity implies that modules themselves are made up by smaller modules. It was introduced by Ravasz and co-workers aiming to resolve the apparent contradiction of modularity and scale-free structure of networks. In the analysis of metabolic networks, they associated a decreasing clustering coefficient for highly connected nodes with hierarchical modular organization (14). In such networks, poorly connected nodes (i.e. the majority of proteins) are situated in modules and thus have a large clustering coefficient. In contrast, hubs connecting these distinct modules display only small clustering coefficients. We observed this dependency of clustering coefficient on degree for most networks compared (fig. 4). For orthology-based networks, however, this pattern is absent suggesting the lack of a hierarchical structure in these networks. Alternatively, large highly connected complexes could result in proteins having both a large number of interactions and large clustering coefficient.

4 Discussion and Conclusions

Graph theory represents an important and popular approach for the analysis of large-scale interaction networks (12). It is frequently used to obtain a general characterization of molecular networks. It is also of importance for systems biology in revealing modular structures which can subsequently be modeled more quantitatively. Nevertheless, results of graph-theoretical analyses should be taken with caution, since they are generally based on the assumption that the studied network is error-free and complete. This, however, is hardly the case for current protein interaction networks. Whereas the effects of sparse sampling have been intensively studied, the impact of the method used to generate interaction networks has been neglected so far (22,23).

Here we presented a first graph-theoretical comparison of major human interaction networks. It shows that the method used for the network assembly strongly influence the structure of the network. While all interaction networks showed small world properties and corresponded to scale-free networks, we detected considerable structural divergence regarding their modularity and hierarchical structure. This observation has to be taken into account for an unbiased application of graph theory. General conclusions about the structure of the human protein interaction network should therefore be verified against potential interference by the chosen assembly method. As many previous analyses of network structures are based on single networks, a reassessment of their results might be warranted.

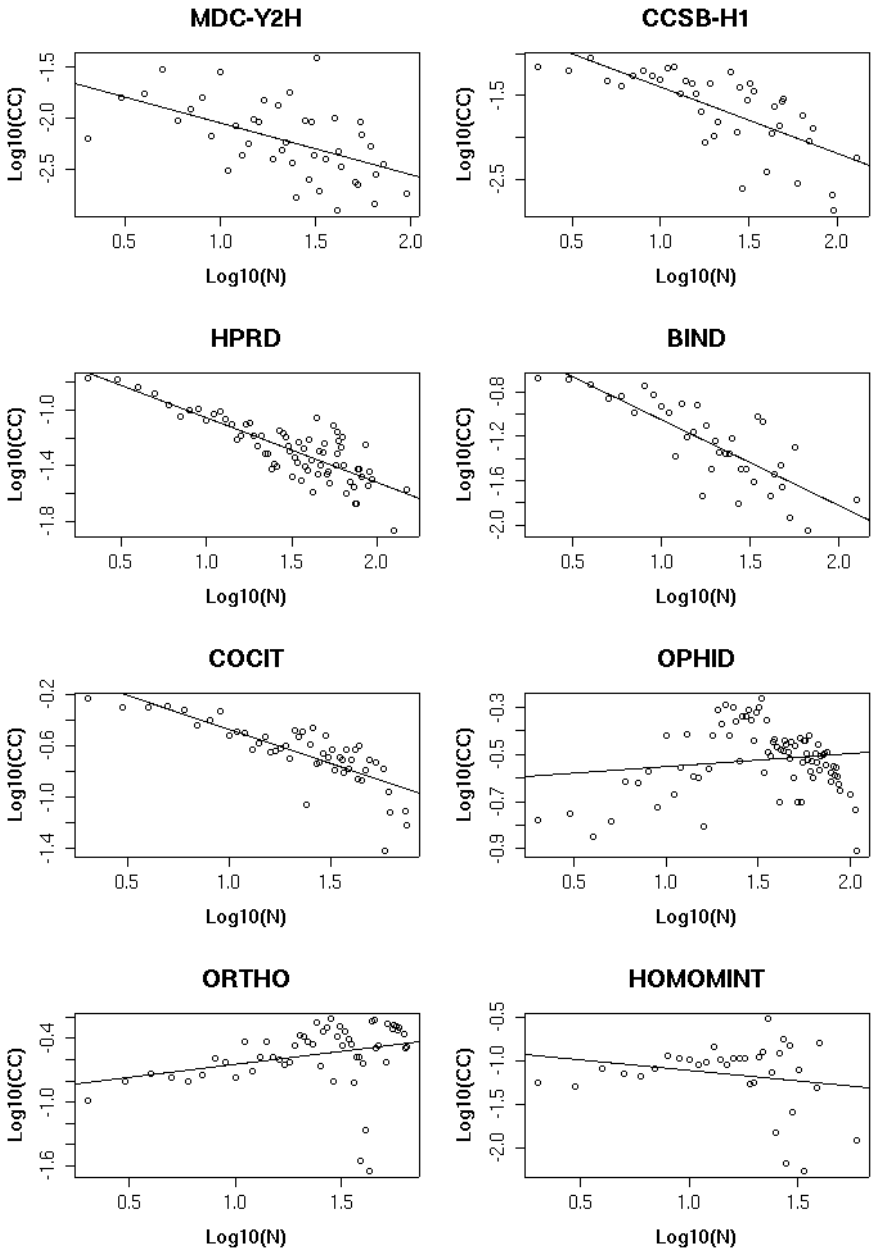

Figure 4: Clustering coefficient. Plots show the dependence of the clustering coefficient on the degree of proteins. The clustering coefficients shown were derived by averaging over all proteins having the same degree. The solid line shows the linear fit.

Supplementary materials

Supplementary materials can be found at: http://itb.biologie.hu-berlin.de/Members/ futschik/ ibsb2007

Acknowledgements

The work presented here was supported by the SFB 618 grant of the *Deutsche* Forschungsgesellschaft (DFG). We thank Bronwyn Carlisle for careful proofreading.

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. Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5,** 101-113.
13. Lim, J. *et al.* (2006) A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, **125,** 801-814.

14. Ravasz, E. *et al.* (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297,** 1551-1555.
15. Wagner, A. (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, **18,** 1283-1292.
16. Ramani, A.K. *et al.* (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, **6,** R40.
17. Chaurasia, G. *et al.* (2007) UniHI: an entry gate to the human protein interactome. *Nucleic Acids Res*, **35,** D590-594.
18. Dijkstra, E.W. (1959) A note on two problems in connection with graphs. *Numerische Mathematik*, **1,** 269-271.
19. Carey, V.J. *et al.* (2005) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21,** 135-136.
20. Balasubramanian, R. *et al.* (2004) A graph-theoretic approach to testing associations between disparate sources of functional genomics data. *Bioinformatics*, **20,** 3353-3362.
21. Lee, I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science*, **306,** 1555-1558.
22. Han, J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430,** 88-93.
23. de Silva, E. *et al.* (2006) The effects of incomplete protein interaction data on structural and evolutionary inferences. *BMC Biol*, **4,** 39.