Systematic Functional Assessment of Human Protein-Protein Interaction Maps

Gautam Chaurasia[1,2]
g.chaurasia@biologie.hu-berlin.de

Hanspeter Herzel[1]
h.herzel@biologie.hu-berlin.de

Erich E. Wanker[2]
ewanker@mdc-berlin.de

Matthias E. Futschik[1]
m.futschik@biologie.hu-berlin.de

[1] Institute for Theoretical Biology, Humboldt-Universität, Berlin, Germany
[2] Max-Delbrück-Centrum for Molecular Medicine, Berlin-Buch, Berlin, Germany

Abstract

Protein-protein interaction maps can contribute substantially to the discovery of protein cooperation patterns in the cell. Recently, several large-scale human protein-protein interaction maps have been generated using experimental or computational approaches. Evaluation of these maps is likely to provide a better understanding of human biology. However, careful analysis is needed, as the comparison of interaction maps of lower eukaryotes showed a surprising divergence between different maps. Here, we present a first systematic functional assessment of eight currently available large-scale human protein-protein interaction maps. The analysis shows that these maps include a large number of common proteins, but only a small number of common interactions. We detected several types of biases that need to be considered in the future utilization of these maps.

Keywords: protein-protein interaction, human interactome, networks, systems biology

1 Introduction

Protein-protein interaction (PPIs) networks underlie most processes in a cell, and play a crucial role in the formation of structural complexes, intracellular signaling, cell-cell communication and virtually every other aspect of cellular function. Comprehensive analysis of PPIs is therefore essential for the understanding of how proteins cooperate in the cell. In recent years, we have witnessed large-scale protein interaction mapping projects in several organisms such as *S. cerevisiae*, *D. melanogaster* and *C. elegans*. Now, the focus has moved towards the systematic mapping of human protein-protein interactions [2, 3, 8, 10, 11, 12, 14, 16]. These human PPI maps have been derived from both computational and experimental approaches and are likely to play an important role in biology and medicine. However, studies of interaction maps in lower eukaryotes showed a surprisingly small degree of overlap between different maps [1, 17, 19]. Thus, critical comparison of the available human interaction maps is necessary regarding the congruency of different data sources. Whereas such comparisons have been performed for yeast, they are still lacking for human PPI maps.

We conducted a first systematic functional assessment of eight recently published large-scale human PPI maps [2, 3, 8, 10, 11, 12, 14, 16]. We started our analysis by examining the overall number of common proteins and interactions in the analyzed maps. To investigate their composition with regard to protein function, we utilized the Gene Ontology (GO) annotation database to determine the distribution of the functional classes covered by proteins from each map. We subsequently assessed the functional coherency of interaction maps using two different approaches. Finally, a hierarchical cluster analysis was performed to delineate the relationship between different biological processes based on interactions listed in maps [15, 19].

Table 1: Summary of eight different large-scale human interaction maps. Here, **P** is the actual number of proteins, and **PM** denotes the number of proteins mapped to EntrezGene ID. Similarly, **I** is the actual number of interactions, whereas **IM** represents the number of interactions after mapping.

Network	P	PM	I	IM	Methods	Reference
MDC-Y2H	1703	1703	3186	3186	Y2H-assay	Stelzl *et al.* 2005, *Cell*
CCSB-H1	1549	1549	2754	2754	Y2H-assay	Rual *et al.* 2005, *Nature*
HPRD	6206	5908	20940	15658	Literature-review	Peri *et al.* 2003, *Gen. Res.*
BIND	4275	2677	5872	4233	Literature-review	Bader *et al.* 2001, *NAR*
COCIT	3737	3737	6580	6580	Literature-mining	Ramani *et al.* 2004, *Gen. Biol.*
OPHID	4787	2284	24993	8962	Homology-based	Brown *et al.* 2005, *Bioinformatics*
SANGER	3870	3503	11651	9641	Homology-based	Lehner *et al.* 2003, *Gen. Bio.*
HOMOMINT	4129	2556	10182	5582	Homology-based	Persico *et al.* 2005, *BMC Bioinformatics*

2 Materials

2.1 Assembly of Human Protein-Protein Interaction Maps

Currently large-scale PPI maps are mainly based on either yeast-two-hybrid (Y2H) assays, literature review or homology. In the following, we give a short description of each approach and the interaction maps included in this analysis:

(i) **Y2H-assays** are based on *in vivo* binding of fused proteins: To detect interaction between two proteins, the first protein is fused to a DNA binding domain of a transcription factor, its potential binding partner to the corresponding activation domain of the transcription factor. Any interaction between them is reported by subsequent formation of an intact and functional transcriptional activator [5]. Using this technology, two large-scale screenings for human PPIs were performed by Stelzl *et al.* [16] and Rual *et al.* [14]. These maps were included in our analysis and referred to as MDC-Y2H and CCSB-H1.

(ii) In **literature-based approaches**, information about PPIs is cataloged by scientists through reviewing published literature. Two representative databases are the Human Protein Reference Database (HPRD) [10] and the Biomolecular Interaction Network Database (BIND) [2]. For our comparison, we extracted all binary PPIs stored in these databases and constructed corresponding interaction maps. Alternatively to manual literature search, Ramani and co-workers used computational text-mining algorithms to gain information about PPI from Pubmed abstracts [12]. This computationally generated interaction map is referred to as COCIT in this analysis.

(iii) **Homology-based** interaction maps are derived from interactions observed in other organisms. This approach assumes that interactions are conserved for orthologous proteins [9]. Thus, human PPI maps can be computationally predicted exploiting observed PPIs in model organisms such as *S. cerevisiae*, *D. melanogaster* and *C. elegans*. The first homology-based human PPI map was constructed by Fraser and Lehner [8] using the InParanoid algorithm to determine orthologous human proteins [13]. It will be referenced as SANGER map in our analysis. We decided to include only the so-called SANGER core data set, as it contains high-confidence protein interactions. We also included two further homology-based mappings that are currently stored in the Online Predicted Human Interaction Database (OPHID) [3] and the HOMOMINT database [11].

Altogether, a total of ∼30,000 proteins and ∼86,000 interactions were collected from these eight different sources. Since these PPIs were generated under different conditions using different identifiers, a crucial task was to unify the data under a shared naming and annotation convention. For this purpose, each protein was mapped to its corresponding EntrezGene ID. A small loss of proteins and interactions was seen in the conversion of BIND, HPRD and SANGER maps, whereas more than ∼50%

of the proteins and interactions were lost in the case of OPHID and HOMOMINT respectively, mainly due to lack of EntrezGene ID. Table 1 summarizes the number of proteins and interactions before and after mapping. Overall, we could map ~25,000 proteins and ~57,000 interactions from the eight networks, of these ~10,700 proteins and ~52,000 interactions were unique. These were the basis for our further analysis.

For the efficiency of computational analysis, all interaction maps were converted into graph objects using the *graph* package from Bioconductor [6].

3 Results

3.1 Common Proteins and Interactions in Human PPI Maps

For yeast, comparisons showed strikingly small overlap between different large-scale interaction maps [17, 19]. Therefore, we examined whether this is also the case for the human PPI maps. We first determined the number of common proteins and interactions. The results are presented in Figure 1. Whereas a majority of proteins (60%) was found in multiple interaction maps, only a small percentage of interactions (~8%) were supported by more than one method. The small overlap between PPI maps is possibly due to the high rate of both false positives and negatives.

Not a single interaction was found in six or more networks. Only 8 interaction pairs (ARCN1-COPB, FBP1-FBP1, MAX-MYC, MXI1-MAX, PDHA1-PDHB, MAPK3-MAP2K1, PTS-PTS, LSM2-LSM3) appeared in five different networks. Interestingly, most of these pairs are either homodimers or formed by homologous proteins. As six of the eight pairs were supported by at least two homology-based interaction maps, one may speculate that these interactions resulted from gene duplication during evolution [18].

Regarding common proteins, we found only eleven proteins (CKS2, MAPK14, EEF1G, NFKBIA, PARK2, PIN1, PEX5, QARS, TP53, TTR) in all eight networks. Remarkably, all of these proteins are highly connected. For example, EEF1G (elongation factor 1 gamma) and TP53 (tumor protein) have 359 and 315 unique interacting partners, respectively. The remaining proteins (CKS2, MAPK14, NFKBIA, PARK2, PIN1, PEX5, TTR) have between 100 and 130 interaction partners. This is especially noteworthy since the average number of interaction partners per protein is rather small, ranging from 1.8 (for CCSB-H1) to 3.9 (for OPHID).

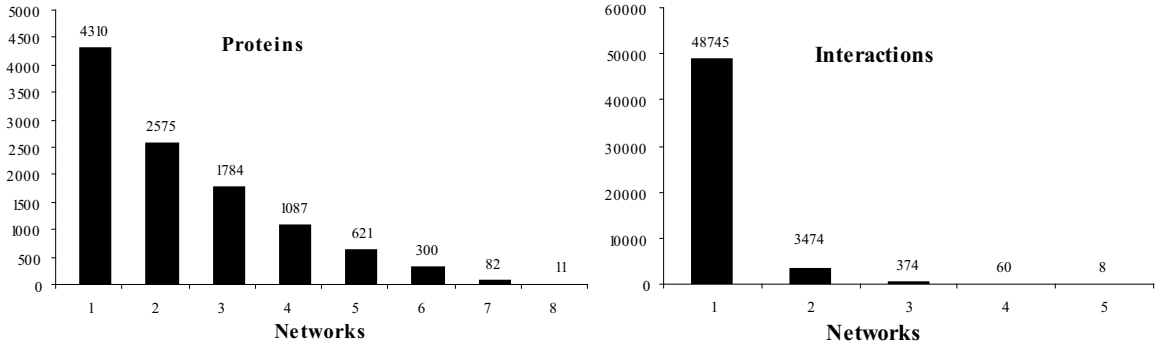

Figure 1: Frequency of common proteins and interactions in PPI maps compared. The x-axis refers to the number of networks in which common proteins and interactions were found.

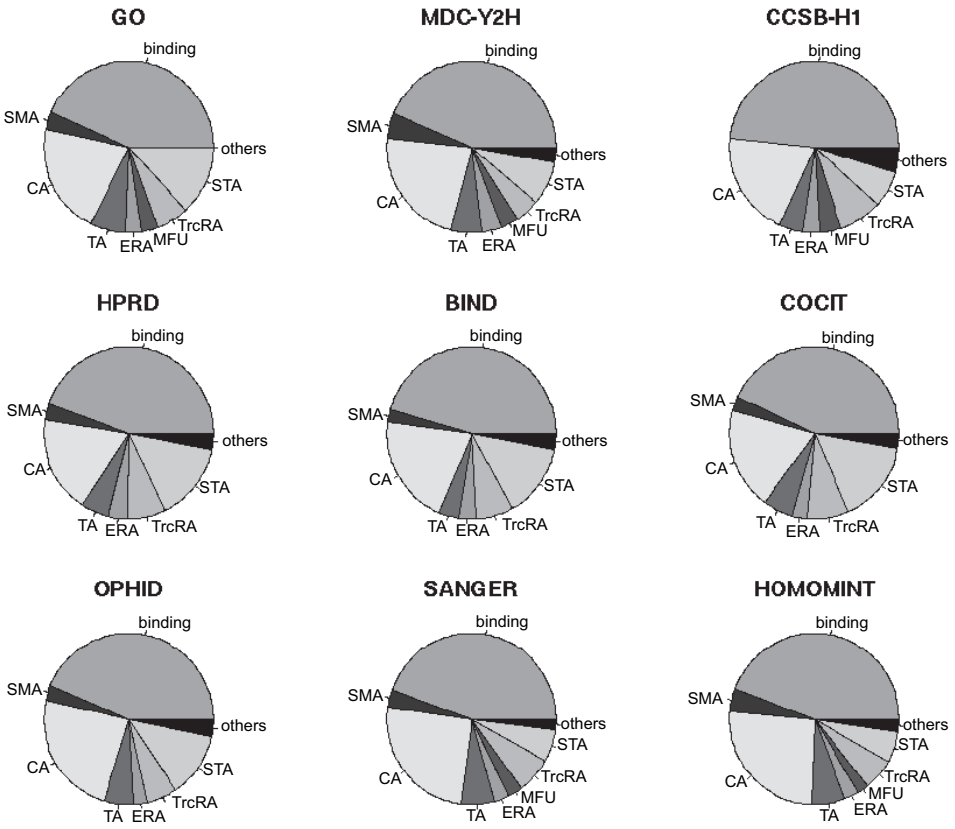

Figure 2: Distribution of the functional categories covered by proteins of the maps compared. The composition of all human genes annotated in GO is shown as reference. The following abbreviation were used to represent the functional classes: **SMA** - Structural Molecule Activity; **CA** - Catalytic Activity; **TA** - Transporter Activity; **CRA** - Chaperone Regulator Activity; **ERA** - Enzyme Regulator Activity; **MFU** - Molecular Function Unknown; **TrcRA** - Transcription Regulator Activity; **TCAAAA** - Triplet Codon Amino Acid Adaptor Activity; **TrnRA** - Translation Regulator Activity; **NRA** - Nutrient Reservoir Activity; **PT** - Protein Tag; **STA** - Signal Transducer Activity; **Others** - $< 3\%$ (AA, CRA).

3.2 Functional Annotation of Network Proteins

Besides small overlap, previous assessments of yeast protein interaction maps have revealed that most methods for generating interaction maps have their own characteristic biases. Especially, their composition regarding functional classes is influenced by the mapping procedure chosen [17, 19]. We carried out an analysis similar to the mentioned yeast studies, in order to detect potential sampling and selection biases in the human PPI maps. To determine the functional composition of maps compared, proteins were classified based on their annotation in Gene Ontology (GO) [7]. In specific, we analyzed the maps in regard to molecular function categories at the first level of the GO hierarchy. Figure 2 shows the distribution of the functional categories covered by different maps. At first glance, the observed distribution of functional classes looks similar for all eight maps. The largest category is 'binding' for all maps, representing ~50% of proteins included. This is to be expected in an analysis of interaction maps. The second largest functional category in all maps (~29–35%) is 'catalytic activity' including kinases known to be major regulators of cellular pathways. Literature-based maps show an over-representation of signaling proteins (~20–24%) as compared to homology-based and Y2H-based

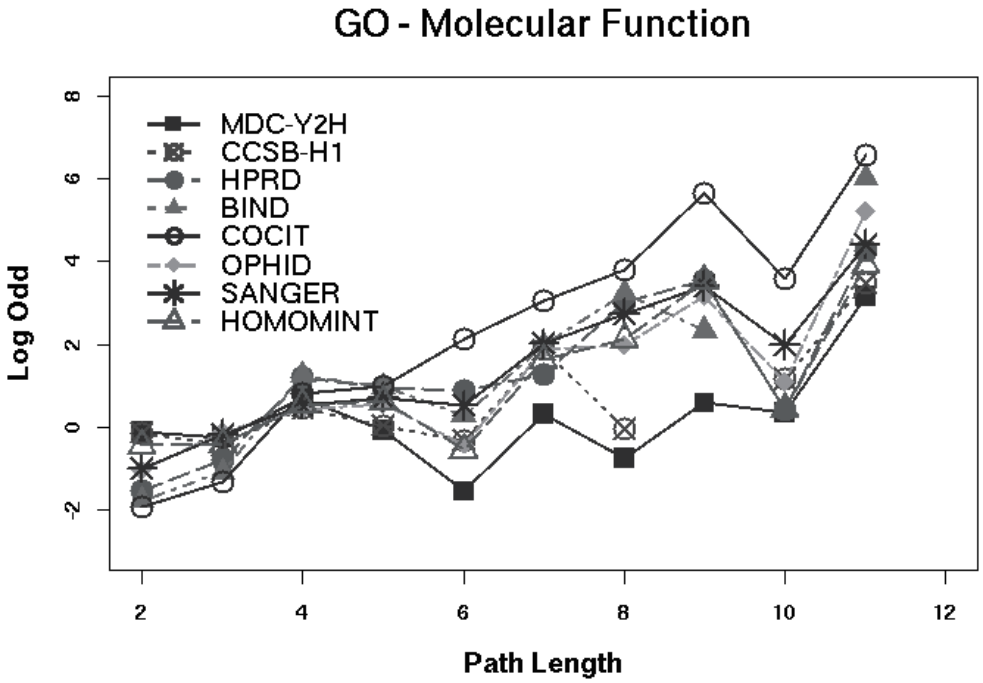

Figure 3: Assessment of the functional coherency of all eight maps. The similarity scores were calculated using shared path length of two similar GO terms. Path-length distribution of the original network was compared to that of the randomized networks. Log-odd scores were calculated to measure the difference between original network and a random network, and plotted against the path length.

maps. Nearly ∼5% proteins were assigned to the category 'molecular function unknown' (MFU) from Y2H-based and homology-based (SANGER, HOMMINT) maps. Thus, these maps might be especially useful for *de novo* functional annotation of hitherto uncharacterized proteins.

3.3 Assessment of Functional Coherency

To assess the reliability of interaction data sets, it has recently been suggested to examine the similarity of GO terms assigned to interacting proteins [8, 11]. The reasoning behind this approach is that interacting proteins frequently share function. Therefore, we expect from reliable PPI maps that proteins of related function are more likely to interact than proteins of unrelated function. We used increased interaction probability of functionally related proteins as a first benchmark to scrutinize the accuracy of human PPI maps. For each pair of proteins, we estimated the similarity of the corresponding GO annotation for the category molecular function. The relatedness of two GO terms was determined by measuring their proximity in the GO hierarchy. Starting from the root term, related GO terms are expected to have larger shared path length than unrelated GO terms. For each pair of proteins, we determined the corresponding pair of GO terms and then computed their shared path length. To test the significance, we generated random networks containing the same number of nodes connected by the same number of edges. For any pair of interacting proteins, in which both proteins have GO annotation, we computed the difference (log-odd scores) between the path length in the original network and in the random network. Finally, log-odd scores were plotted against the shared path length (Figure 3). In the case of strong correlation between function and interaction, we expect higher log-odd score for larger shared path length. It is clear from the figure that COCIT data shows the strongest correlation between function and interaction, whereas HPRD, BIND, OPHID,

SANGER, and HOMOMINT exhibit only moderate correlation. Weaker correlation is observed in both Y2H-based maps, and a clear correlation is visible only for the path length 11.

3.4 Shared GO Term Similarity Measurement

An alternative approach to assess the quality of interaction data sets is to calculate the frequency of protein interactions between identical functional categories [1, 17]. This defines an interaction matrix $I_{(m,n)}$, where, m and n refer to functional classes. Similarly, we expect that reliable interaction data sets show large diagonal elements in the matrix I, indicating the over-representation of interactions between proteins assigned to related functional categories. We applied this as a second benchmark to human PPI maps. Each pair of proteins was assigned to their functional classes using molecular categories at the third level of GO hierarchy [7]. For these categories, we calculated the interaction matrix I. To assess the significance, we computed the log-odd scores to measure the deviation of the observed frequency distribution k_{mn} with an expected base line distribution $\mathrm{k}^0 mn$. Results from this analysis are summarized in Figure 4. The patterns observed varied between different maps. Large log-odd scores along the diagonal of the interaction matrix were found for all literature-based interaction maps. Similarly, homology-based maps except OPHID also displayed these patterns. This indicates functional coherency of interactions in these maps. Interestingly, the log-scores for some categories were independent of the map chosen (e.g. cation-binding). This observation may point to particularly strong binding for such categories.

3.5 Hierarchical Clustering of Biological Processes

In the previous sections we examined the quality of the interaction data sets by analyzing protein interactions mainly between proteins of the same functional category. Proteins in a network, however, do not only interact within the same functional class but also across classes [15, 19]. The interactions between categories can be characterized on the basis of the topology of PPI maps, using a hierarchical clustering approach. The number of links between proteins from two different functional classes is used to measure their association. Here, we employed this strategy to delineate the relationship between proteins involved in different biological processes (BP). The proteins of each pair were assigned to their biological processes based on GO tree at level 3. For determining the degree to which proteins from two different classes are connected, we first computed the ratio matrix between original and random network, defined as follows

$$R_{(m,n)} = \frac{O_{(m,n)}}{P_{(m,n)}}$$

where $R_{(m,n)}$ is the ratio matrix, $O_{(m,n)}$ and $P_{(m,n)}$ are matrices containing the total number of links between biological processes m and n in an original and in a random network respectively. A distance matrix was then computed using this ratio matrix.

$$D_{(m,n)} = \frac{1}{R_{(m,n)}}$$

For relating biological processes, we used hierarchical clustering based on the average linkage-clustering algorithm. The algorithm uses the distance matrix $D_{(m,n)}$, and places those BP classes close to each other that are topologically closely related, i.e. have many interactions.

As an example, we looked in detail at the clusters obtained for the BIND interaction map (Figure 5). We observed the 12 different biological processes to be grouped in two main clusters. Cluster one can be further divided in two sub-clusters. The larger sub-cluster includes all cellular regulatory processes, signaling transduction and macromolecule metabolism. These biological processes appear to be tightly coupled and to construct a densely connected network, shown by strong enrichment of interactions. It is also interesting to see that the biological process 'cell cycle' cluster with 'negative regulation

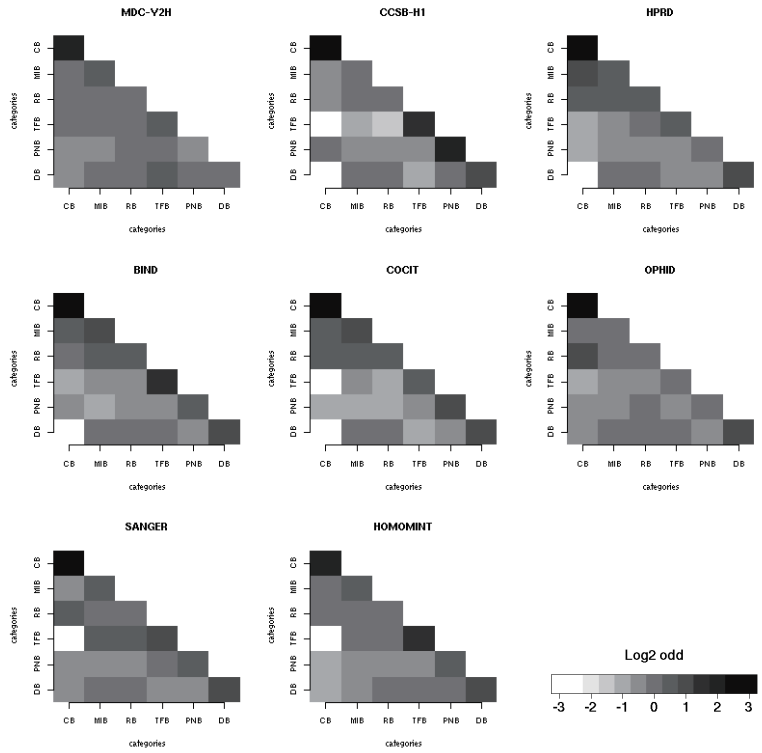

Figure 4: Functional annotation of the interaction matrix $I_{(m,n)}$ representing the intensity distribution of interaction between functional classes m and n for all interaction maps. The following abbreviations were used to represent the categories along x and y axes: **CB** (cation binding), **MIB** (metal ion binding), **RB** (receptor binding), **TFB** (transcription receptor binding), **PNB** (purine nucleotide binding), **DB** (DNA binding). Each pair of proteins was assigned to their functional classes using molecular functional categories at third level 3 of GO hierarchy. Log-odds scores were computed to measure the difference between original network and those obtained for random network, and used to display matrices. Interactions between same categories are represented along the diagonal of the matrix, whereas interaction between different categories are ordered from top to bottom as rows and from left to right as column. For example matrix element top left corner shows the overrepresentation of self-interaction between cation-binding. As a second example, matrix element $I_{(3,1)}$ shows interaction between cation-binding and transcriptor factor binding, which is underrepresented.

of cellular processes' and 'regulation of metabolism'. These findings indicate that the cell cycle is tightly controlled with negative regulation. The smaller sub-cluster consists of primary and cellular metabolism which shows a weak correlation of intensity distribution (light-blue). The second main cluster contains the biological processes involved in 'response to stress and external stimulus'. This group is also characterized by a weak coupling.

Clusters obtained for HPRD and COCIT interaction maps also show a similar pattern as for BIND interaction map (see online supplementary information Figure 3 and Figure 4). In contrast, homology-based interaction maps display different patterns with primary, cellular and macromolecule metabolism being densely connected. This group closely interacts with another group containing cell organization and biogenesis, cell cycle and regulation of metabolism (see online supplementary information Figure 5, Figure 6 and Figure 7). Finally, clustering patterns for Y2H-based interaction map CCSB-H1 resemble those obtained for literature-based maps, whereas MDC-Y2H shows similar clustering pattern as for homology-based interaction maps (see online supplementary information Figure 1 and Figure 2).

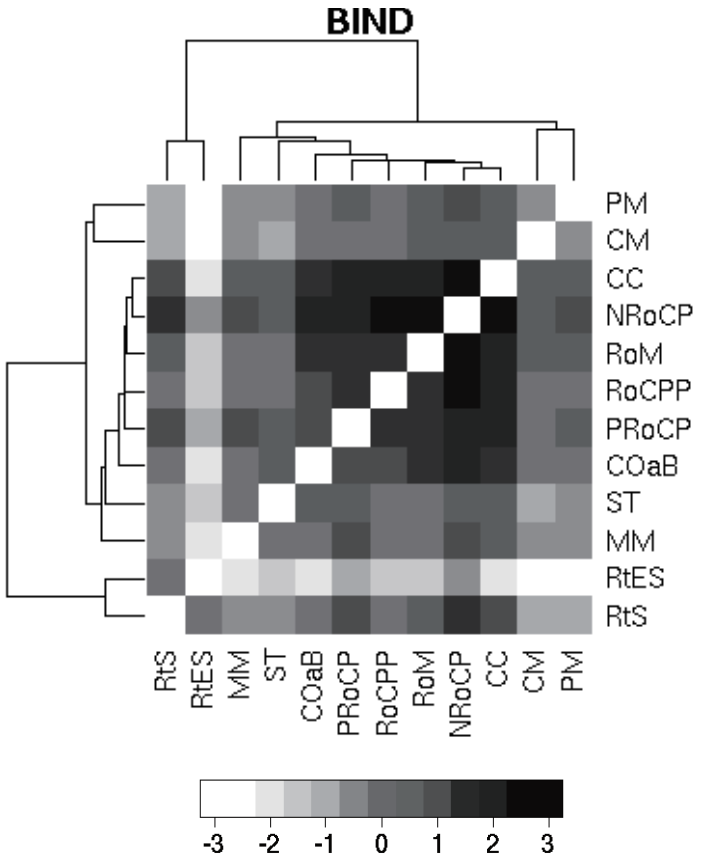

Figure 5: Hierarchical clustering of the biological processes based on GO annotation database for BIND interaction map. Diagonal of the matrix shows the interaction between same categories and is set to zero. Positive intensity distributions are the overrepresentation of the interaction between different categories. **ST**- signal transduction, **PRoCP** - positive regulation of cellular process, **NroCP** - negative regulation of cellular process, **RoCPP** - regulation of cellular physiological process, **CC** - cell cycle, **CM** - cellular metabolism, **COaB** - cell organization and biogenesis, **RoM** - regulation of metabolism, **RtS** - response to stress, **RtES** - response to external stimulus, **PM** - primary metabolism, **MM** - macromolecule metabolism.

4 Conclusions

We have presented a first systematic functional assessment of eight different human protein-protein interaction maps. These maps were derived either from Y2H-assays, literature reviews or homology-based on interactions between orthologous proteins in other organisms. The results from the distribution of common proteins and common interactions show that current maps exhibit only small overlap between networks. The majority of proteins can be seen in multiple maps, but only ∼8% of the interactions are found in more than one network.

As reported in previous studies of yeast protein interaction maps, most approaches for generating PPI maps have different tendencies and selection biases. We detected strong sampling and detection biases linked to the method of map generation. It is noticeable from the results that the distributions of functional classes, obtained for eight different maps, agree on some properties of PPI maps. Indeed, we found that in all interaction maps binding proteins were over-represented, followed by proteins involved in catalytic activities. A significant over-representation of signaling proteins was observed in all literature-based maps, as compared to other homology-based and Y2H-based maps.

This distribution of functional classes is explained by the fact that many signaling pathways have been studied and published in the past at low-throughput experimental scale. This has resulted in the enrichment of signaling proteins in literature-based maps. However, we also found that the functional class 'membrane proteins' is under-represented in all maps.

We applied two different approaches for assessing the accuracy of maps. Both approaches were based on the observation that the interaction of proteins is likely to correlate with their functional properties, and thus to similar GO annotation. In all analyses, functional coherency varied between different maps. Literature-based interaction maps showed stronger correlation between interaction and function than homology-based and Y2H-based interaction datasets. These results indicate that literature-based maps offer more reliable data sets. However, the quality of literature-based maps might be overestimated as GO is also based on published literature and therefore, does not provide a fully independent benchmark. In contrast, more novelty can be expected from the Y2H-based maps, since these maps are not based on as many *a priori* assumptions.

We also illustrated how protein interaction maps can be used to unravel the complex relationships between different biological processes. For future systematic analysis, we are currently implementing a database, called UniHi [4], which is designed to integrate various large-scale human PPI maps. The objective of this database is to provide a convenient platform to support scientists undertaking large-scale system biology.

References

[1] Bader G. D. and Hogue C. W., Analyzing yeast protein-protein interaction data from different sources, *Nat. Biotechnol.*, 20(10):991–997, 2002.

[2] Bader G. D., Betel D., and Hogue C. W., BIND: The biomolecular interaction network database, *Nucleic Acids Res.*, 31(1):248–250, 2003.

[3] Brown K. R. and Jurisica I., Online predicted human interaction database, *Bioinformatics*, 21(9):2076–2082, 2005.

[4] Chaurasia, G. *et al.*, UniHI: An entry gate to the human protein interactome, *in preparation.*

[5] Fields S. and Song O., A novel genetic system to detect protein-protein interactions, *Nature*, 340(6230):245–246, 1989.

[6] Gentleman R. C., Carey V. J., Bates D. M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y., Gentry J., Hornik K., Hothorn T., Huber W., Iacus S., Irizarry R., Leisch F., Li C., Maechler M., Rossini A. J., Sawitzki G., Smith C., Smyth G., Tierney L., Yang J. Y., and Zhang J., Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biol.*, 5(10):R80, 2004.

[7] Harris M. A., Clark J., Ireland A., Lomax J., Ashburner M., Foulger R., Eilbeck K., Lewis S., Marshall B., Mungall C., Richter J., Rubin G. M., Blake J.A., Bult C.,Dolan M., *et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, 32:D258–D261, 2004.

[8] Lehner B. and Fraser A. G., A first-draft human protein-interaction map, *Genome Biol.*, 5(9):R63, 2004.

[9] Matthews L. R., Vaglio P., Reboul J., Ge H., Davis B. P., Garrels J., Vincent S., and Vidal M., Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or interologs, *Genome Res.*, 11(12):2120–2126, 2001.

[10] Peri S., Navarro J. D., Amanchy R., Kristiansen T. Z., Jonnalagadda C. K., Surendranath V., Niranjan V., Muthusamy B., Gandhi T. K., Gronborg M., Ibarrola N., Deshpande N., Shanker K., Shivashankar H. N., Rashmi B. P., *et al.*, Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res.*, 13(10):2363–2371, 2003.

[11] Persico M., Ceol A., Gavrila C., Hoffmann R., Florio A., and Cesareni G., HomoMINT: An inferred human network based on orthology mapping of protein interactions discovered in model organisms, *BMC Bioinformatics*, 6 Suppl 4:S21, 2005.

[12] Ramani A. K., Bunescu R. C., Mooney R. J., and Marcotte E. M., Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome, *Genome Biol.*, 6(5):R40, 2005.

[13] Remm M., Storm C. E., and Sonnhammer E. L., Automatic clustering of orthologs and in-paralogs from pair wisespecies comparisons, *J. Mol. Biol.*, 314(5):1041–1052, 2001.

[14] Rual J. F., Venkatesan K., Hao T., Hirozane-Kishikawa T., Dricot A., Li N., Berriz G. F., Gibbons F. D., Dreze M., Ayivi-Guedehoussou N., Klitgord N., Simon C., Boxem M., Milstein S., Rosenberg J., *et al.*, Towards a proteome-scale map of the human protein-protein interaction network, *Nature*, 437(7062):1173–1178, 2005.

[15] Schwikowski B., Uetz P., and Fields S., A network of protein-protein interactions in yeast, *Nat. Biotechnol.*, 18(12):1257–1261, 2000.

[16] Stelzl U., Worm U., Lalowski M., Haenig C., Brembeck F. H., Goehler H., Stroedicke M., Zenkner M., Schoenherr A., Koeppen S., Timm J., Mintzlaff S., Abraham C., Bock N., Kietzmann S., Goedde A., Toksoz E., Droege A., Krobitsch S., Korn B., Birchmeier W., Lehrach H., and Wanker E. E., A human protein-protein interaction network: A resource for annotating the proteome, *Cell*, 122(6):957–968, 2005.

[17] von Mering C., Krause R., Snel B., Cornell M., Oliver S. G., Fields S., and Bork P., Comparative assessment of the large-scale data sets of protein-protein interactions, *Nature*, 417(6887):399–403, 2002.

[18] Wagner A., The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.*, 18(7):1283–1292, 2001.

[19] Yook S. H., Oltvai Z. N., and Barabasi A. L., Functional and topological characterization of protein interaction networks, *Proteomics*, 4(4):928–942, 2004.