RESEARCH ARTICLE

Quantitative assessment of the structural bias in protein–protein interaction assays

Åsa K. Björklund[1], Sara Light[2], Linnea Hedin[1] and Arne Elofsson[1]

[1] Department of Biochemistry and Biophysics, Center for Biological Membrane Research/Stockholm Bioinformatics Center, Stockholm University, Stockholm, Sweden
[2] Science and Technology Computing Division, Computation Directorate, Lawrence Livermore National Laboratory, Livermore, USA

Received: February 15, 2008
Revised: May 23, 2008
Accepted: June 25, 2008

With recent publications of several large-scale protein–protein interaction (PPI) studies, the realization of the full yeast interaction network is getting closer. Here, we have analysed several yeast protein interaction datasets to understand their strengths and weaknesses. In particular, we investigate the effect of experimental biases on some of the protein properties suggested to be enriched in highly connected proteins. Finally, we use support vector machines (SVM) to assess the contribution of these properties to protein interactivity. We find that protein abundance is the most important factor for detecting interactions in tandem affinity purifications (TAP), while it is of less importance for Yeast Two Hybrid (Y2H) screens. Consequently, sequence conservation and/or essentiality of hubs may be related to their high abundance. Further, proteins with disordered structure are over-represented in Y2H screens and in one, but not the other, large-scale TAP assay. Hence, disordered regions may be important both in transient interactions and interactions in complexes. Finally, a few domain families seem to be responsible for a large part of all interactions. Most importantly, we show that there are method-specific biases in PPI experiments. Thus, care should be taken before drawing strong conclusions based on a single dataset.

Keywords:
Abundance / Disorder / Protein–protein interactions / Tandem affinity purification / Yeast two hybrid

1 Introduction

In the last decade, knowledge about the functions of many proteins has been gathered through high-throughput protein–protein interaction (PPI) experiments [1–10]. Clearly, the use of large-scale assays to understand the functions of proteins has been very useful. However, the number of overlapping interactions between the different studies is surprisingly small, mainly due to the high rate of false negatives in PPI assays, *i.e.* that the current techniques cannot detect a large portion of all true interactions [11–14]. In addition, false positives may also be a cause of small overlaps.

For example, consider the two independent genome-wide screens for the complexes of the *Saccharomyces cerevisiae* proteome published in 2006 [5, 6]. Both used tandem affinity purification (TAP) where a bait protein is expressed with a dual tag for purification in two steps, first on IgG columns and subsequently on Calmodulin columns. Finally, the purified preys are identified using MS. The two steps of purification should ensure as few contaminants as possible.

Correspondence: Professor Arne Elofsson, Department of Biochemistry and Biophysics, Stockholm University, SE-106 91 Stockholm, Sweden
E-mail: arne@bioinfo.se
Fax: +46-8-164672

Abbreviations: GO, gene ontology; **LCI,** literature curated interactions (refers to the dataset by Reguly *et al.*); **Pfam,** the protein family database; **PPI,** protein–protein interactions; **SVM,** support vector machines, **TAP,** tandem affinity purification; **Y2H,** yeast two hybrid screen

In the end, however, a comparison between the identified protein complexes of the two near-identical studies, showed substantial differences [12].

Another widely used method for identifying PPI is the yeast two hybrid (Y2H) screen [15]. As opposed to TAP, that identifies protein complexes, Y2H identifies direct pairwise interactions between two proteins. The method is based on the reconstitution of a transcription factor and thereby expression of a reporter protein: the DNA-binding domain and the activation domain of the factor are fused to two proteins of interest. Due to the differences in methodology, the overlap between Y2H screens and TAP experiments is, as expected, quite low [11, 14].

Aside from the interest in deciphering the function of individual proteins, large-scale interactions have served to identify general properties that are of importance for interaction participation. For instance, it has been proposed that proteins with disordered regions [16–19] and high content of repeating domains [18, 20] are important for interactions, in particular transient ones. These proteins contain large surfaces for flexible binding, hence they may accommodate many interactions *in vivo*. An alternative view is that these properties make the proteins 'sticky' so that they interact with practically anything and therefore cause many false positives in interaction assays. Furthermore, experiments indicate that highly interactive proteins (hubs) often are well conserved and/or essential [21, 22]. In addition, homologous proteins, and in particular proteins with domains from the same family, tend to interact more frequently than others [23, 24]. Finally, proteins that interact are often coexpressed.

However, some of these features have only been shown for a single dataset and may not be as pronounced in other datasets. Therefore, we investigate the differences between the various *S. cerevisiae* PPI datasets and assess their reliability. We aim to characterize the inherent biases in the different methods. Further, we want to see whether these biases can explain part of the small overlap between different interaction assays.

2 Methods

2.1 Protein–protein interaction datasets

The interaction data from two TAP studies, Gavin *et al.* [5] and Krogan *et al.* [6], was extracted using the spoke model [25]. In the spoke model, there are links between the tagged protein (bait) and each protein in its harvested complex (preys), but there are no links between preys in the same complex. This is in contrast to the matrix model [25] which is used for the MIPS datasets [26], where links exist between all members of the same complex. The MIPS database is divided into different levels of complexes and we only considered interactions between proteins that belong to the same complex at the lowest level. Y2H data were downloaded from IntAct [27], where all interactions from two hybrid experiments, as well as two hybrid fragment pooling, were included. The Reguly data were downloaded from the Journal of Biology website in August 2006 [28].

Due to the differences in size, it is not straightforward to compare the Gavin and Krogan datasets. In order to get a better comparison of the TAP datasets, we created a reduced Krogan set (Krogan-R), of similar size as the Gavin set. This was done by extracting high confidence interactions only, with LC-MS/MS confidence score over 99.6 or a MALDI score over 2 as defined in [6].

2.2 Protein properties

Cutoffs for the different properties mentioned below, were selected in such a manner that, when possible, there were a similar number of proteins with each property. That is, around 900–1000 proteins with the property as can be seen in Table 2. The protein family database (Pfam)-A [29] domains were assigned using HMMER (http://hmmer.wustl.edu) with a cutoff for assignments at an E-value of 0.1. Within domain repeats, additional assignments with higher E-values were allowed as previously described [20]. In addition, related domains were grouped together according to the Pfam Clans (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/). Proteins were considered repeating if two or more domains from the same family/clan is adjacent to each other.

Promiscuous domains were defined as the domains that occur most frequently in interactions. In other words, the sum of the connectivity of all proteins containing the domain, taking all five datasets into account. Alternative cutoffs for promiscuous domains were tested (Supporting Information). The ten most promiscuous domains were Actin-like ATPase domain, Domain present in DEAD box RNA helicases (DEAD), G-protein superfamily, β propeller, Helicase conserved C-terminal domain (Helicase C), RNA motif recognition domain (RRM), Armadillo/β-catenin-like repeat (ARM), ATPase family associated with various cellular activities (AAA), NADP-binding Rossman fold and Protein Kinase domain.

Disordered regions were predicted using DISOPRED 2.1 [30]. A protein was considered disordered if 80 or more contingent residues with predicted disorder are found. Conserved proteins were defined as proteins from a KOG [31] that is present in six eukaryotes. A list of essential yeast proteins were defined according to a study by Giaever *et al.* [32]. Abundant proteins were defined as those having more than 6000 molecules *per* cell according to ref. [33]. Membrane proteins were defined as those having more than one predicted transmembrane helix by TMHMM [34].

2.3 Functional assignment

Functional annotations from Gene Ontology (GO) were downloaded from the *Saccharomyces* Genome Database (ftp://ftp.yeastgenome.org/yeast/) and only terms at the more general level of GO slim were considered.

2.4 SVM predictions

Support vector machine (SVM) training and testing was performed using the SVM-light package [35]. A simple classification scheme was used, where the features for each protein pair were represented with the numbers 0, 1 or 2. Each property was divided in a binary mode with presence or absence of the property. As an example, if the property is repeated/not repeated: 0 = none of the two proteins are repeated, 1 = one of them is and 2 = both are. SVMs with a polynomial kernel were trained to predict interacting proteins with each of the datasets as positive examples and an equal number of random protein pairs as negative ones. In each round of classification, one-third of the dataset was kept for testing while training was performed on the remaining two-thirds. Ten rounds of training and testing were run to calculate SDs.

2.5 Statistical tests

To estimate the statistical significance of the results, *Z*-scores were calculated from randomization in 10 000 iterations. The *Z*-score was calculated as $Z = |x - \mu|/\sigma$ where x is the observed value and μ is the average value obtained from simulations with SD σ. Assuming a normal distribution of the data, the *p*-value was then derived from the *Z*-score using normal distribution *p*-value tables.

For the enrichment of properties in the different datasets random pairs were drawn from a set of proteins with the same fraction of that property as is found in the whole yeast proteome.

3 Results and discussion

3.1 Interaction datasets

We have investigated three experimental yeast interaction datasets referred to as Krogan *et al.* [6], Gavin *et al.* [5], and Y2H [15]. In addition, we have included two manually curated datasets for quality assessment. These are MIPS complexes [26] and a literature curated set by Reguly *et al.* [28], referred to as LCI. The latter is comprised of PPIs gathered from a literature survey of small-scale experiments [28]. In addition, a third curated dataset from Kiemer *et al.* [36] was also investigated with similar results as the LCI dataset (Supporting Information). This dataset contains small-scale interactions detected with at least two independent methods.

The Krogan and Gavin datasets were obtained from large-scale TAP experiments and, like the MIPS database, describe protein complexes rather than pairwise interactions. It is noteworthy that not all members of a complex are necessarily in direct physical contact. Also, we have used the spokes method [25] to define the interactions in the TAP datasets (Section 2) and hence, there may be a bias towards interactions involving proteins that were successfully cloned as TAP baits. In addition, the purification steps should disable detection of transient interactions [3]. In many respects, Y2H detection is complementary to TAP experiments in that it only detects direct, physical pairwise interactions, and may detect weak, transient interactions, that are missed with TAP.

Despite two purification steps, there seem to be many false positives in the TAP experiments [5, 6]. Therefore, filtration algorithms have been applied on the raw data in both studies to obtain biologically meaningful interactions. Here, however, we have performed our analysis primarily on the raw data since we are interested in the interactions each method can detect, rather than analyzing complexes that may differ due to the filtering process. Indeed, the interactions that many may regard as false positives may also tell us about the biases of different methods.

The perhaps most striking difference between the two datasets is the large number of interactions in the Krogan set compared to the other datasets, see Table 1. Therefore, a reduced Krogan set (Krogan-R), containing a similar number of interactions as Gavin, was constructed (Section 2). The number of baits used in the Gavin and Krogan experiments is similar and more than 50% of the baits are identical. Interestingly, each bait has about twice as many identified preys in the Krogan dataset compared to Gavin. The discrepancy can be explained by different detection methods; while MALDI-TOF was used for prey identification in both experiments, Krogan *et al.* also used a more sensitive detection through LC-MS/MS.

The number of pairwise interactions that are shared between the three experimental datasets can be seen in Fig. 1. The intersection between the Gavin and Krogan datasets is small with only 4048 common pairwise interactions. As may be expected, due to the differences in methodology, the intersection between Y2H data and the two TAP datasets is

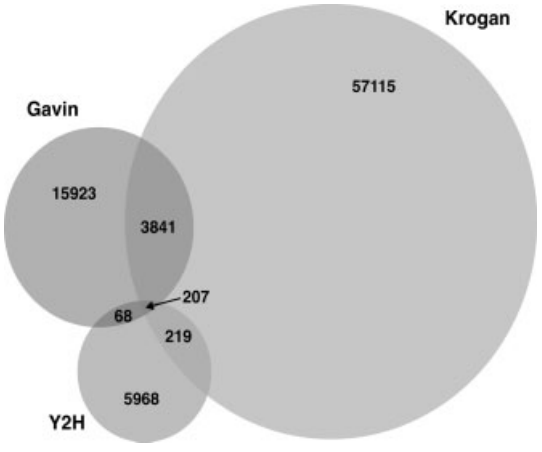

Figure 1. Overlap between datasets. The overlap between pairwise interactions in Gavin, Krogan and Y2H datasets. The numbers indicate the number of protein pairs in each field.

Table 1. Datasets

Dataset	noP	noPairwise	k/P	noBait	k/Bait	noPrey	k/Prey
Gavin	2576	20039	7.78	1759	12.02	1814	11.7
Krogan	5359	61382	11.45	2264	27.98	5321	11.9
Krogan-R	2792	20353	7.29	2064	10.34	2425	8.8
Y2H	3863	6462	1.67	–	–	–	–
LCI	3307	11858	3.59	–	–	–	–
Mips	1073	9869	9.20	–	–	–	–

The number of proteins (noP) and pairwise interactions (noPairwise) followed by the average connectivity (k/P) in each dataset. For Gavin and Krogan TAP data, the number of baits (noBait) and preys (noPrey) is shown followed by the average number of preys *per* bait (k/Bait) and baits *per* prey (k/Prey). The Krogan dataset consists of preys identified through MALDI-TOF (2363) and LC-MS/MS (5201).

extremely low, but still highly significant compared to random sampling from all possible interactions (*Z*-scores: 114 for Gavin and 97 for Krogan).

3.2 Assessing the quality of the interactions

The GO provides a controlled vocabulary for genes [37]. It consists of three classifications: molecular function, cellular compartment and biological process. Although it is not evident that a pair of interacting proteins should have the same molecular function, they should primarily be colocalized and participate in the same biological process. Hence, GO annotations may serve to roughly estimate the reliability of the interaction datasets.

Here, we have used a version of GO that was retrieved before the release of the two TAP studies [5, 6], and it is therefore safe to assume that we are not, in a circular manner, using functional annotations derived from these datasets for assessment. However, the same cannot be said in the case for the Y2H dataset which thus may have an overestimated GO overlap.

As can be seen in Fig. 2, a large fraction of the interacting pairs from all datasets are annotated to the same compartment and process using GO slim [37]. The fraction of colocalized interacting proteins is similar in the Gavin and Krogan datasets, although it is substantially lower than the corresponding fraction in the curated LCI dataset (Fig. 2). The fraction of interacting pairs involved in the same biological process is highly significant in all datasets (Supporting Information), but it is higher in the Gavin dataset compared to Krogan unless the Krogan-R is considered. Notably, the intersection of Gavin and Krogan (denoted G + K in Fig. 2) is enriched in protein pairs with the same GO annotations and reaches a level similar to what is seen for LCI. This suggests, as could be expected, that the interactions found in both of the studies are of higher confidence.

Another measure of quality is the overlap with literature curated interactions in LCI and MIPS. Although these datasets should contain less false positives compared to high-throughput experiments, the small-scale interaction experiments contained therein suffer from inspection bias, *i.e.* involve proteins of particular interest to the scientists performing the experiments. A relatively small fraction of the Krogan data is found in MIPS and LCI, while the overlap with the Gavin dataset is larger (Fig. 2). However, if the Krogan-R is considered, the overlap with curated interactions is similar to that of Gavin. As with GO, the intersection between Gavin and Krogan is clearly enriched in more reliable interactions. Unfortunately, the Y2H set cannot be evaluated in a similar manner since LCI and our Y2H dataset contain interactions from the same source. Instead, we compared it to the Kiemer dataset [36] that contains small-scale interactions detected by two independent methods (Supporting Information). However, this comparison gave nearly identical results as the LCI dataset, probably due to the same problem as LCI, an overlap with the Y2H set.

In conclusion, using these measures, Gavin seems to be of slightly better quality than the full Krogan set, but the datasets are of similar quality when reduced to a similar size. Further, the intersection between Gavin and Krogan are clearly of better quality, while it is hard to draw any conclusions about the Y2H dataset from these results due to biases in the evaluation methods.

3.3 Protein properties and interactions

The structural properties of a protein affect its propensity to form interactions with other proteins and the nature of those interactions. For instance, several studies show that proteins with high connectivity, *i.e.* hub proteins, contain more repeated protein domains and disordered regions [16–18, 38, 39]. Given the disparate experimental procedures generating the PPI datasets, it is possible that certain protein properties are particularly enriched in some datasets. Here, we have studied the bias of the datasets towards structural properties and other characteristics associated with connectivity.

Most other investigations of PPI networks tend to focus on hubs as compared to nonhubs. However, the definition of a hub is often an arbitrary cutoff at a certain level of connectivity. We have chosen, instead, to take into account all the

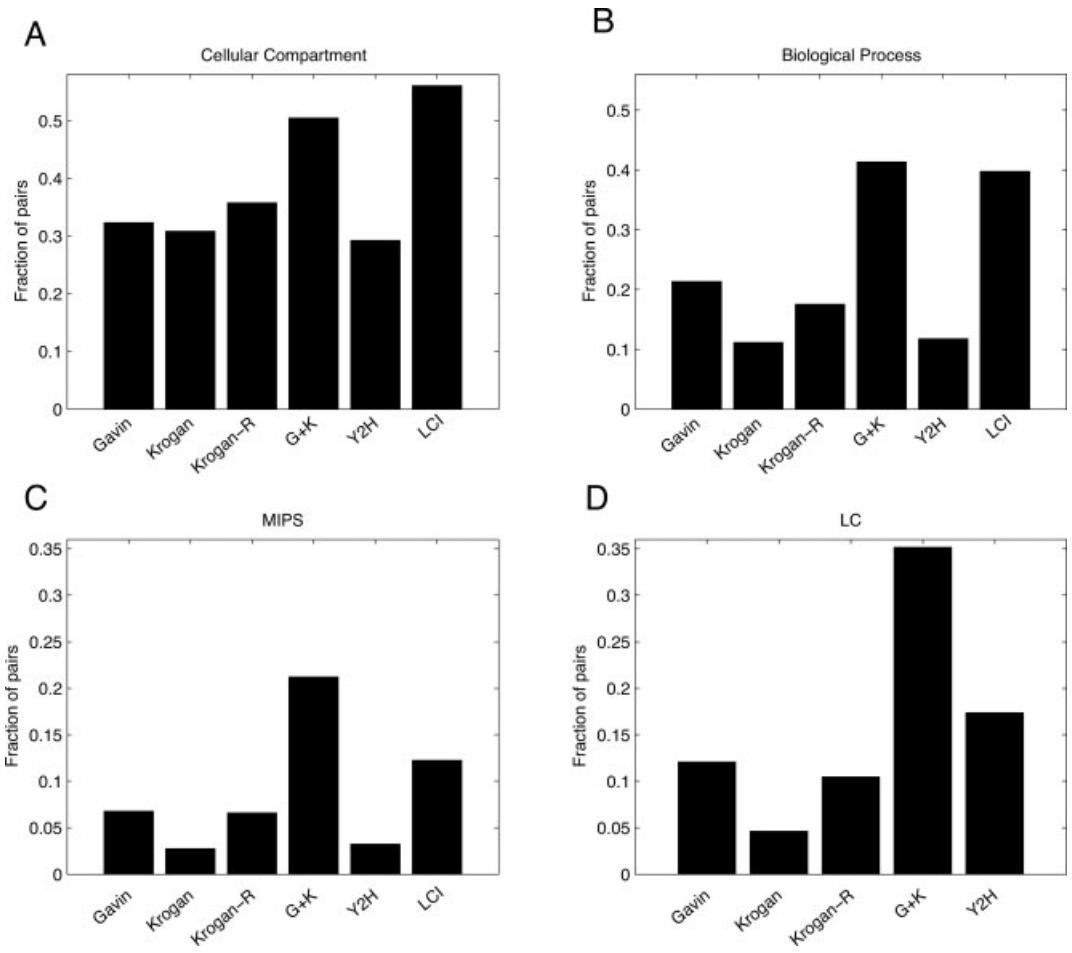

Figure 2. Quality of experimental datasets. Assessment of interactions by GO annotations and literature curated datasets. The fraction of interacting protein pairs that have the same GO slim annotation for cellular compartment (A) and biological process (B) followed by the fraction of interactions that are found in MIPS (C) and LCI (D). The datasets shown are Gavin, Krogan, Krogan-R, Gavin–Krogan intersection (G + K), Y2H and LCI.

pairwise interactions and try to avoid classification of the proteins into different categories. Hence, we have investigated how frequently different protein properties are found in one or both proteins in an interacting pair.

3.3.1 Disordered regions

Disordered regions are not associated with any tertiary structure, but instead allow the protein to assume different states. These regions enable flexibility and are of particular importance for molecular recognition [30, 40]. Therefore, it is not surprising that disordered regions are often found in hub proteins [16–19, 39] and are likely to be enriched in transient interactions [18, 19].

The enrichment of interacting protein pairs involving a disordered protein is shown in Fig. 3A. It is clear that there are substantial differences between the datasets with respect to disorder. The Y2H and Krogan datasets show considerable over-representation of interactions involving a disordered protein. However, this trend cannot be seen in the Gavin dataset. This difference between Gavin and Krogan does not seem to be related to the detection methods as we find that the Krogan interactions detected using LC-MS/MS, as well as by MALDI-TOF, show over-representation of disordered proteins (Supporting Information). Therefore, the lack of disordered proteins in the Gavin dataset does not appear to be due to the general principle of the pull-down method, nor due to the detection method, but is due to some as yet unidentified factor.

Further, there is no significant enrichment of disorder in the Gavin–Krogan intersection, indicating that many of the interactions with disordered regions may be considered less reliable. Alternatively, the explanation lies in that disorder proteins are generally associated with low abundance (Table 2), a feature which decreases the likelihood of detection through several pull-down methods. A confounding factor is the strong over-representation of disorder/protein interactions in the LCI dataset (Supporting

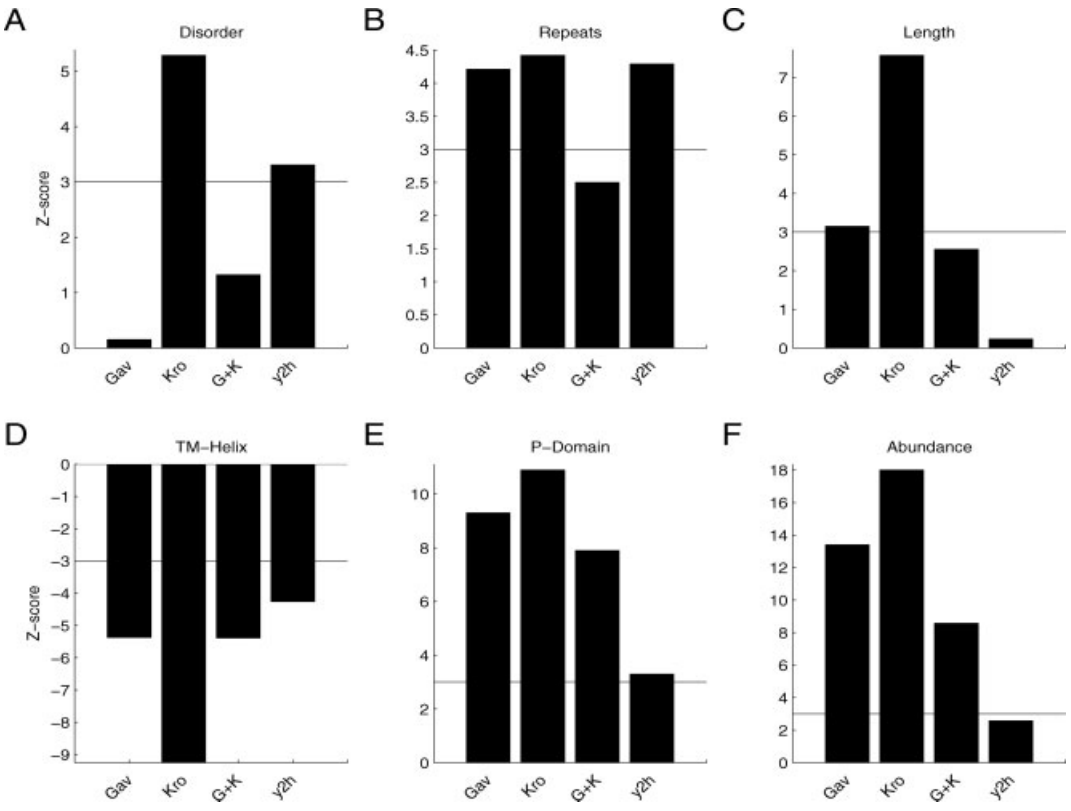

Figure 3. Enrichment of physical protein properties. *Z*-scores for the fraction of PPI pairs that contain a protein that (A) contains disordered regions >80 residues; (B) has a domain repeat; (C) is longer than 800 residues; (D) has more than one predicted transmembrane helix or (E) has a promiscuous domain or (F) is abundant (over 6000 molecules/cell). The datasets shown are Gavin (Gav), Krogan (Kro), Gavin–Krogan intersection (G + K) and Y2H. The *Z*-scores are calculated as $Z = |x - \mu|/\sigma$, where x is observed fraction, μ the expected fraction from randomization with SD σ. A *Z*-score of three corresponds to a *p*-value of 0.0001.

Table 2. Correlation between properties

Dataset	Proteins	Overlap with properties, number of proteins (*Z*-score)							
		Domain repeat	Disorder	Length	P-domain	TMH	Abundance	Conserved	Essential
Domain repeat	487	0(0.0)	137(10.5)	151(13.5)	158(17.3)	50(−1.1)	98(4.7)	188(12.8)	129(7.8)
Disorder	971	137(10.5)	0(0.0)	404(31.1)	244(17.6)	74(−4.0)	93(−3.2)	284(11.2)	159(2.3)
Length	907	151(13.5)	404(31.1)	0(0.0)	230(16.8)	99(−0.6)	119(0.3)	228(7.2)	198(7.3)
P-domain	742	158(17.3)	244(17.6)	230(16.8)	0(0.0)	23(−7.6)	192(11.1)	292(17.1)	198(10.4)
TMH	885	50(−1.1)	74(−4.0)	99(−0.6)	23(−7.6)	0(0.0)	57(−6.1)	129(−1.9)	74(−5.2)
Abundance	1002	98(4.7)	93(−3.2)	119(0.3)	192(11.1)	57(−6.1)	0(0.0)	397(20.9)	281(14.2)
Conserved	1342	188(12.8)	284(11.2)	228(7.2)	292(17.1)	129(−1.9)	397(20.9)	0(0.0)	499(28.0)
Essential	110	129(7.8)	159(2.3)	198(7.3)	198(10.4)	74(−5.2)	281(14.2)	499(28.0)	0(0.0)

The total number of yeast proteins with each property followed by the co-occurrence with other properties in a protein. The *Z*-scores are in parenthesis; for over(positive)/under(negative)-representation of the co-occurrence of two properties.

Information). There, protein interactions detected through pull-down methods are only marginally less likely to contain disorder proteins compared to interactions derived from Y2H experiments and other miscellaneous biochemical methods (data not shown).

In conclusion, compared to all other methods we have examined, the Gavin experiment protocol disfavours interactions involving proteins with disordered regions. That aside, our results indicate that disorder is over-represented both in the binary interactions detected by Y2H and in TAP complexes.

3.3.2 Repeats of protein domains

With their highly variable sequences and frequent copy variation, repeated domains create large surfaces that can interact with a variety of partners [20, 41, 42]. Domains such as the tetracopeptide repeat can be found with different numbers of repeating units in yeast and each of these repeats interact with several different interaction partners [41]. We have previously shown that domain repeats are over-represented in eukaryotic interaction networks [18, 20]. In addition, sequence repeats, including repeated domains as well as shorter repeats, are common in hub proteins [38].

Here, a protein with domain repeats is defined as a protein with one or more tandem Pfam [29] domains from the same family. We find that they are strongly enriched in pairwise interactions in all datasets (Fig. 3B). They seem to be equally important in interactions found with TAP or Y2H. However, in the intersection of Gavin and Krogan (G + K), probably the most reliable subset of the high-throughput methods, this trend is weaker, possibly indicating that domain repeats can give rise to false positives.

3.3.3 Protein size

Protein size, here defined as the length of the protein sequence, is clearly correlated to disordered regions and domain repeats (Table 2). To accommodate long domain repeats and large disordered regions, a protein needs to be of substantial size. Further, large proteins can provide more binding surfaces, and could therefore be expected to interact with many partners. Hence, it is not clear if the sheer size of the proteins leads to the observed over-representation of disordered regions and domain repeats in interactions.

As can be seen in Fig. 3C, proteins longer than 800 residues are over-represented in the two TAP datasets, especially Krogan. However, the distribution of long proteins does not show the same patterns as disorder or domain repeats (Figs. 3A and B), and especially the enrichment of repeats and disorder in Y2H is clearly not only due to their size as long proteins are not enriched in the Y2H set. In addition, Huang *et al*. [14] conducted a study of the false discovery rate in Y2H experiments and found that it was not higher for large protein than small, hence at least in Y2H, there should be no methodological bias towards larger proteins.

3.3.4 Membrane proteins

Membrane proteins are not well represented in the PPI datasets due to the experimental obstacles involved in detecting their interactions. Indeed, we find that while the expected fraction of membrane proteins among the interacting pairs of proteins is around 22%, most datasets contain about 5% membrane protein interactions. Hence, there is a significant under-representation of membrane proteins in the PPI datasets (Fig. 3D). Furthermore, the Y2H dataset contains almost twice as many membrane protein interactions as any other dataset, which is somewhat surprising considering that these interactions have been detected in the nucleus [15] using constructs that are unlikely to detect genuine membrane protein interactions. It has also been shown by Huang *et al*. [14] that membrane proteins can give rise to many false positives in Y2H screens.

3.3.5 Promiscuous domains

The domains that a protein is composed of are the building blocks that provide binding surfaces between proteins. The protein properties discussed above may be related to the domain content of the protein. Nevertheless, interactions between certain domain families may be preferred over others. In addition, some domains may have a more pronounced tendency to interact with many different partners. Here, we refer to such domains as 'Promiscuous domains'.

As seen in Fig. 4, for a selection of domain families, the same domains are often over-represented in Gavin and Krogan. This implies that, while the number of overlapping interactions is small, they involve similar domain combinations. For instance, the low overlap may be explained by natural variations in expression patterns of similar proteins. The domains that are frequently found in the TAP data are mainly those that occur often in the proteome. Still, some less common domains are over-represented in the interactions, in particular, the superfamily Actin-like ATPase. The high frequency of this domain family can primarily be explained by a few yeast chaperons from the HSP70 family with very high connectivities in both TAP datasets. It is questionable whether these interactions should be considered false positives, as they are probably not part of the final functional complex, however, they are likely to occur *in vivo*. While the two TAP studies contain interactions between common domain families, the Y2H dataset clearly differs from the TAP data with regards to the domains found in the interactions. Here, instead, the Cyclin, SH3 and Pumilio-family RNA-binding repeat domains are over-represented.

Without structural mapping of the interaction interfaces, we cannot specify the interacting domains. Nevertheless, we have approximated the promiscuous domains as the ones found in most interacting proteins. Thus, a universal set of promiscuous domains was defined from all interactions in our five datasets. As expected, these include many typical protein-binding domains such as the AAA ATPase, the actin ATPase and ARM. However, many nucleotide-binding domains such as DEAD, Helicase C and RRM are also found in the dataset. This finding could indicate that the domains are involved both in nucleotide-binding and protein interactions, that the interaction is mediated through nucleotide binding, or simply that proteins with these domains have many interaction partners through other domains in the proteins. A comprehensive list of the promiscuous domains and their interactions can be found in the Supporting Information. The ten most promiscuous domains (Section 2) are also quite common and occur in 11% of all yeast proteins.

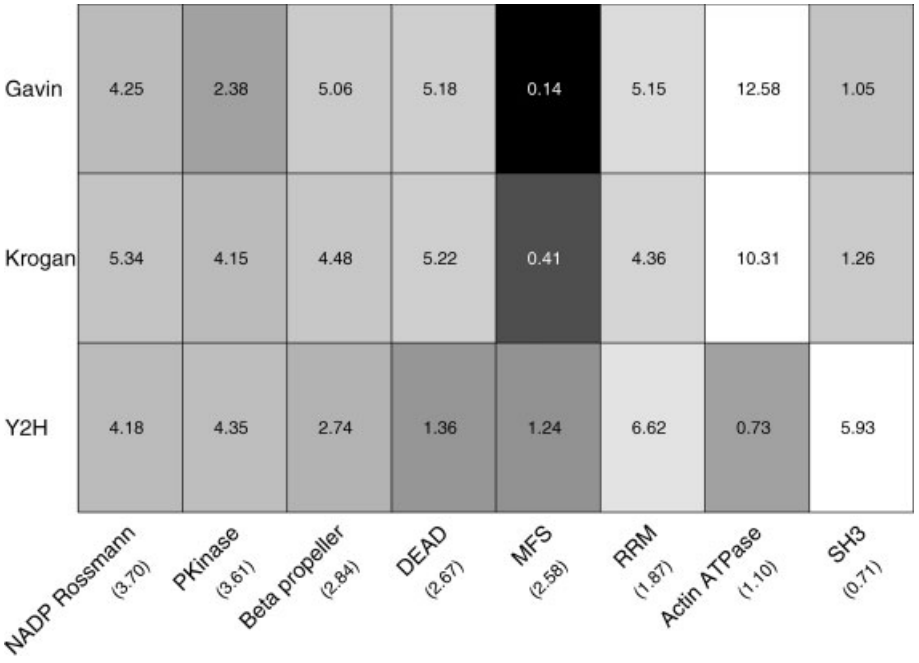

Figure 4. Domains in interactions. For a selection of domain families the over(light)/under (dark)-representation in each interaction dataset is shown. The number below each name is the expected fraction of protein pairs with that domain while the number in each box is the observed value in each dataset (NADP Rossmann = NADP-binding Rossmann fold, Pkinase = protein kinase domain, β propeller, DEAD = domain present in DEAD box RNA helicases, MFS = major facilitator superfamily (12–14 TM-region permease), RRM = RNA motif recognition domain, Actin ATPase = actin-like ATPase domain, SH3 = Src homology domain.

All datasets show a strong statistical over-representation of interactions involving these ten domains, indicating that a few promiscuous domains are responsible for a large fraction of all interactions (Fig. 3E). In fact, nearly half of all interactions include one or more of these ten domains. Alternative cutoffs for promiscuous domains were tested (Supporting Information), generating similar results. It should also be noted that proteins with these promiscuous domains also tend to be large and contain domain repeats and disordered structures (Table 2).

3.3.6 Protein abundance

Aside from structural properties, it is well known that abundant proteins tend to be highly connected [11]. Naturally, the more abundant a protein is, the more likely it is to be detected in interaction assays such as TAP. Our results indicate that protein abundance is the strongest contributor to PPI in all of the datasets except Y2H (Fig. 3F). Since, proteins are overexpressed in Y2H there should be no bias towards abundant proteins in that dataset. However, there is still a slight over-representation, possibly owing to the fact that abundant proteins have been assayed more frequently. An alternative, and more interesting, explanation is that abundant proteins participate in a larger number of interactions.

There is a clear correlation between protein abundance and domain repeats (Table 2). Hence, the high frequency of repeats among TAP interactions could primarily be due to their abundance, while this is not a likely explanation in the case of domain repeats in Y2H interactions. Further, disorder and length are actually negatively correlated with abundance. Still, all three properties are strongly over-represented in the Krogan dataset, hence disorder/length must contribute to the interactions to such an extent that they are detected despite their low abundance.

It is also well established that conserved proteins have a higher connectivity than average [21], which is supported in Krogan and Y2H (Fig. 5A). Likewise, essentiality has also been correlated with the connectivity of the proteins in the network [22], and our results show that this is true for all datasets (Fig. 5B). However, there is a strong correlation between abundance, conservation and essentiality (Table 2). Therefore, it is not self-evident which characteristic is the primary cause for the high interactivity of these proteins. Presumably, an essential and conserved protein is more likely to be important in interaction networks [21, 22]. Still, it is unclear whether an abundant protein participates in more

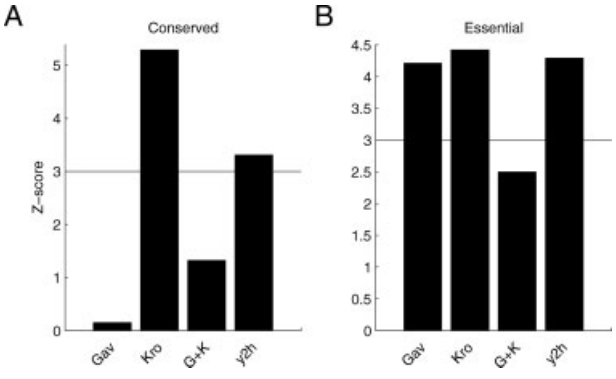

Figure 5. Enrichment of conservation and essentiality. *Z*-scores for the fraction of PPI pairs that contain a protein that (A) is conserved in six species in KOG or (B) is essential. A *Z*-score of three corresponds to a *p*-value of 0.0001.

interactions than its low abundance counterpart. Due to these correlations, here shown to be related to the experimental method, the interactivity of abundant, conserved and essential proteins in the real life network is still unclear.

3.4 Interaction prediction based on physical properties

As of yet, we have identified several properties that are over-represented in interacting protein pairs, and discerned the variation within different datasets (Fig. 3). However, we have also shown that many of these properties are strongly correlated (Table 2). Hence, the relative importance of the investigated properties for the interaction is largely unknown. In an attempt to determine how important the different properties are with regard to interactions, we have performed interaction prediction based on protein properties alone. We have used a SVM that was trained to predict interaction/no interaction-based solely on one or several of the protein properties. If the predictive power is increased by adding more properties this indicates that these properties are contributing independently to the interaction capability.

It is clear from the SVM predictions that abundance is the most important factor for identifying TAP interactions (Fig. 6). Further, although over-represented in some datasets, the individual contributions from domain repeats, disorder and long proteins is low and in some cases close to random (50%). Still, these properties (domain repeats, disorder and length), which are strongly correlated, do provide a small improvement when used in combination in the SVM training, especially in the Krogan set. Nevertheless, abundance gives the best performance in TAP interaction prediction and little is gained by adding the three other properties.

Interestingly, in the Y2H dataset, domain repeats and disorder provide equal classification as abundance, even if it is low. This would suggest that domain repeats and disorder are indeed important for the transient interactions detected by the Y2H method.

4 Concluding remarks

Here, we present an in-depth analysis of the biases in PPI experiments, in particular with regards to a number of physical properties of the interacting proteins. Considering that the TAP and Y2H are quite different methods, it is not surprising that our investigation shows substantial differences with regard to the interacting proteins detected by the two methods. Interestingly, we also demonstrate that the Gavin and Krogan datasets, constructed using similar methods, diverge substantially. In particular, there is a disagreement with regards to disorder content, which is over-represented in Krogan, but not in Gavin. This result may explain why studies based on different datasets have come to conflicting conclusions regarding disorder content in interactions [16–18, 38, 43].

Further, disorder regions should be more common in transient interactions than stable complexes according to recent results [18, 19]. However, the enrichment of disorder in Krogan would suggest that disorder may also be important for interactions in stable complexes, especially if they are detected in spite of their low abundance.

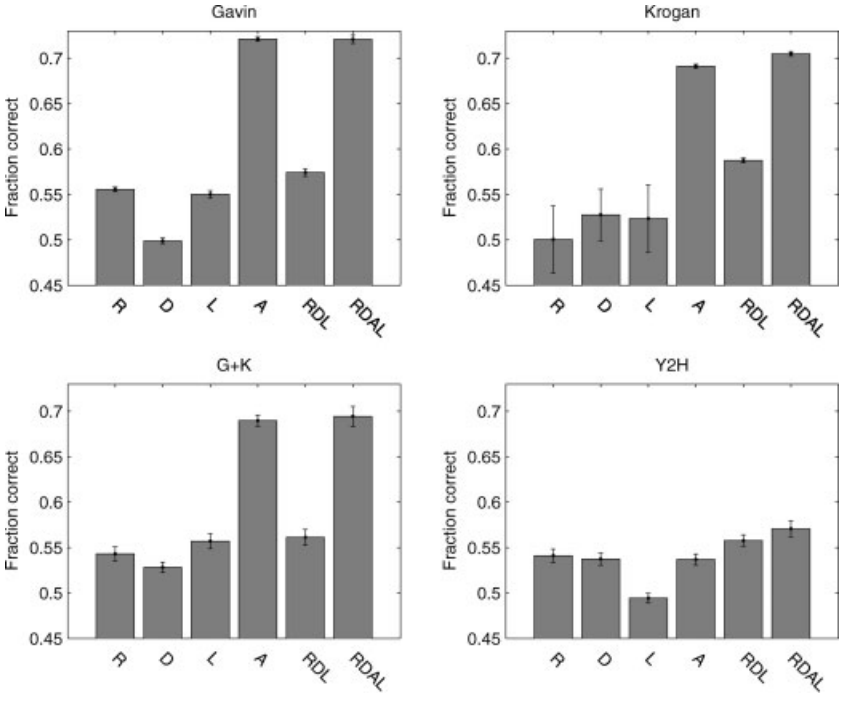

Figure 6. SVM prediction of interactions. The figure shows the percentage correct classified interactions from SVM's trained on different properties of the interacting protein pairs. The properties tested are domain repeats (R), disorder regions >80 residues (D), sequence length >800 residues (L) and protein abundance >6000 molecules/cell (A).

Unlike disordered regions, which are uncommon in the Gavin experiment, repeated domains are common in interactions irrespective of detection method. However, when SVMs were used to discern the best determinant properties for protein interaction, structural properties paled in comparison to the effect of abundance in TAP experiments. This is not surprising, considering that abundant proteins are easier to detect with MS. What remains to be discovered is if abundance is a relevant physical property that is enriched in functional interactions or if it is primarily an artifact of the experimental procedure. In favour of the former is a small enrichment of abundant proteins in Y2H, although abundance should not have any effect on the Y2H assay. The slight over-representation of more abundant proteins in Y2H interactions could indicate that abundant proteins actually have more interaction partners *in vivo*. However, this trend may also be a consequence of sampling biases. Further, a few highly 'promiscuous domains' seem to be involved in nearly half of all interactions. Hence, it would be interesting to further investigate the interactions in which these domains participate, in order to discern if any of these domains give rise to false positives. However, this does not seem to be the case in Y2H interactions as Huang *et al.* [14] do not list these domains as ones responsible for false positives in their study.

In conclusion, long proteins, containing promiscuous domains, disordered structure and domain repeats are prone to interact frequently. However, the degree to which these different properties contribute to interactions will remain an open question until the effect of protein abundance on protein interaction detection is better understood. In the meantime, the best working assumption is that analysis of the full PPI networks should be performed on data stemming from a combination of different methods.

This work was supported by grants from the Swedish Natural Sciences Research Council, SSF (the Foundation for Strategic Research) and the EU Sixth Framework Program is gratefully acknowledged for support to the GeneFun project, contract No: LSHG-CT-2004-503567. Furthermore, this work was partly performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The authors have declared no conflict of interest.

5 References

[1] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 2000, *403*, 623–627.

[2] Ito, T., Chiba, T., Ozawa, R., Yoshida, M. *et al.*, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4277–4278.

[3] Gavin, A. C., Bosche, M., Krause, R., Grandi, P. *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002, *415*, 141–147.

[4] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. *et al.*, Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 2002, *415*, 180–183.

[5] Gavin, A., Aloy, P., Grandi, P., Krause, R. *et al.*, Proteome survey reveals modularity of the yeast cell machinery. *Nature* 2006, *440*, 631–636.

[6] Krogan, N., Cagney, G., Yu, H., Zhong, G. *et al.*, Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* 2006, *440*, 637–643.

[7] Butland, G., Peregrín-Alvarez, J., Li, J., Yang, W. *et al.*, Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* 2005, *433*, 531–537.

[8] Rual, J., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T. *et al.*, Towards a proteome-scale map of the human protein–protein interaction network. *Nature* 2005, *437*, 1173–1178.

[9] Bader, J., Brouwer, C., Chaudhuri, A., Kuang, B. *et al.*, A protein interaction map of *Drosophila melanogaster*. *Science* 2003, *302*, 1727–1736.

[10] Li, S., Armstrong, C., Bertin, N., Ge, H. *et al.*, A map of the interactome network of the metazoan *C. elegans*. *Science* 2004, *303*, 540–543.

[11] von Mering, C., Krause, R., Snel, B., Cornell, M. *et al.*, Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* 2002, *417*, 399–403.

[12] Goll, J., Uetz, P., The elusive yeast interactome. *Genome Biol.* 2006, *7*, 223·1–223·6.

[13] Hart, G., Ramani, A., Marcotte, E., How complete are current yeast and human protein-interaction networks? *Genome Biol.* 2006, *7*, 120·1–120·9.

[14] Huang, H., Jedynak, B., Bader, J., Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* 2007, *3*, e214.

[15] Fields, S., Song, O., A novel genetic system to detect protein–protein interactions. *Nature* 1989, *340*, 245–246.

[16] Dunker, A., Cortese, M., Romero, P., Iakoucheva, L., Uversky, V., Flexible nets: The roles of intrinsic disorder in protein interaction networks. *FEBS J.* 2005, *272*, 5129–5148.

[17] Haynes, C., Oldfield, C., Ji, F., Klitgord, N., Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput. Biol.* 2006, *2*, e100.

[18] Ekman, D., Light, S., Björklund, Å. K., Elofsson, A., What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae? Genome Biol.* 2006, *7*, R45.

[19] Singh, G., Ganapathi, M., Dash, D., Role of intrinsic disorder in transient interactions of hub proteins. *Proteins* 2007, *66*, 761–765.

[20] Björklund, Å. K., Ekman, D., Elofsson, A., Expansion of protein domain repeats. *PLoS Comput. Biol.* 2006, *2*, 959–970.

[21] Eisenberg, E., Levanon, E. Y., Preferential attachment in the protein network evolution. *Phys. Rev. Lett.* 2003, *91*, 128–701.

[22] Jeong, H., Mason, S. P., Barabasi, A. L., Oltvai, Z. N., Lethality and centrality in protein networks. *Nature* 2001, *411*, 41–42.

[23] Pereira-Leal, J., Levy, E., Kamp, C., Teichmann, S. A., Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.* 2007, *91*, 17411433.

[24] Lukatsky, D., Shakhnovich, B., Mintseris, J., Shakhnovich, E., Structural similarity enhances interaction propensity of proteins. *J. Mol. Biol.* 2007, *365*, 1596–1606.

[25] Bader, G., Hogue, C., Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* 2002, *20*, 991–997.

[26] Mewes, H., Amid, C., Arnold, R., Frishman, D., MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* 2004, *32*, D41–D44.

[27] Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., IntAct-open source resource for molecular interaction data. *Nucleic Acids Res.* 2007, *35*, D561–D565.

[28] Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, A. *et al.*, Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* 2006, *5*, 11·1–11·8.

[29] Sonnhammer, E. L., Eddy, S. R., Durbin, R., Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* 1997, *28*, 405–420.

[30] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., Jones, D. T., Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 2004, *337*, 635–645.

[31] Tatusov, R., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 2003, *4*, 41.

[32] Giaever, G., Chu, A., Ni, L., Connelly, C. *et al.*, Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002, *418*, 387–391.

[33] Ghaemmaghami, S., Huh, W., Bower, K., Howson, R. *et al.*, Global analysis of protein expression in yeast. *Nature* 2003, *425*, 737–741.

[34] Sonnhammer, E., von Heijne, G., Krogh, A., A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology., Vol. 6*, AAAI Press, Menlo Park, CA 1998, pp. 175–182.

[35] Joachims, T., Making large-scale SVM learning practical, in Schölkopf, B., Burges, C., Smola, A. (Eds.), *Advances in Kernel Methods—Support Vector Learning*, MIT Press, Cambridge, MA1999, pp. 169–184.

[36] Kiemer, L., Costa, S., Ueffing, M., Cesareni, G., WI-PHI: A weighted yeast interactome enriched for direct physical interactions. *Proteomics* 2007, *7*, 932–943.

[37] Harris, M. A., Clark, J., Ireland, A., Lomax, J. *et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 2004, *32*, D258–D261.

[38] Dosztanyi, Z., Chen, J., Dunker, A., Simon, I., Tompa, P., Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 2006, *5*, 2985–2995.

[39] Patil, A., Nakamura, H., Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.* 2006, *580*, 2041–2045.

[40] Dyson, H., Wright, P., Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 2005, *6*, 197–208.

[41] D'Andrea, L., Regan, L., TPR proteins: The versatile helix. *Trends Biochem. Sci.* 2003, *28*, 655–662.

[42] Binz, H., Amstutz, P., Kohl, A., Stumpp, M. *et al.*, High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* 2004, *22*, 575–582.

[43] Schnell, S., Fortunato, S., Roy, S., Is the intrinsic disorder of proteins the cause of the scale-free architecture of protein–protein interaction networks? *Proteomics* 2007, *7*, 961–964.

