2002–2011 Nucleic Acids Research, 2008, Vol. 36, No. 6 Published online 14 February 2008
doi:10.1093/nar/gkn016

Genome-wide inference of protein interaction sites: lessons from the yeast high-quality negative protein–protein interaction dataset

Jie Guo, Xiaomei Wu, Da-Yong Zhang and Kui Lin*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and College of Life Sciences, Beijing Normal University, Beijing 100875, China

Received September 10, 2007; Revised December 17, 2007; Accepted January 10, 2008

ABSTRACT

High-throughput studies of protein interactions may have produced, experimentally and computationally, the most comprehensive protein–protein interaction datasets in the completely sequenced genomes. It provides us an opportunity on a proteome scale, to discover the underlying protein interaction patterns. Here, we propose an approach to discovering motif pairs at interaction sites (often 3–8 residues) that are essential for understanding protein functions and helpful for the rational design of protein engineering and folding experiments. A gold standard positive (interacting) dataset and a gold standard negative (non-interacting) dataset were mined to infer the interacting motif pairs that are significantly over-represented in the positive dataset compared to the negative dataset. Four negative datasets assembled by different strategies were evaluated and the one with the best performance was used as the gold standard negatives for further analysis. Meanwhile, to assess the efficiency of our method in detecting potential interacting motif pairs, other approaches developed previously were compared, and we found that our method achieved the highest prediction accuracy. In addition, many uncharacterized motif pairs of interest were found to be functional with experimental evidence in other species. This investigation demonstrates the important effects of a high-quality negative dataset on the performance of such statistical inference.

INTRODUCTION

With the advent of high-throughput technologies such as yeast two-hybrid assays (1–5), and the development of various computational methods, either by integrating the vast amount of biological information contained in the genomic datasets (6,7) or by mining from an existing knowledgebase (8,9), rich data resources of interacting proteins have been produced and stored in publicly accessible databases (10–13). Constructing a map of protein–protein interactions is essential not only from a theoretical stance of studying cellular behavior and the machinery of a proteome, but also in the light of potential practical applications such as new drug design (14,15). By intensive analysis and comparison of protein-interaction networks, many studies have emerged to investigate the large-scale biological properties buried in the networks from functional and evolutionary aspects (16), for instance, protein function annotation (17) and interaction interface identification (18). To date, a variety of statistical data analysis techniques have been applied to address these issues, the capability of which depends largely on the accuracy of the protein-interaction dataset (positives), and equally importantly, the non-interaction dataset (negatives).

Currently, high-quality positive datasets have been assembled by combining multiple interaction datasets or integrating additional genomic evidence (19,20). However, the data collected by those methods are far from complete compared with the vast number of possible interactions (21). What makes things more complicated is how to define and assemble a high-quality negative dataset for a statistical analysis system. Negative datasets obviously have a strong effect on the performance of comparative statistical analyses, especially in machine-learning algorithms. The problems induced by lacking negatives cannot be addressed by fine-tuning parameters or finding better statistical methods (22). Currently, two main strategies employed in literatures for assembling negative examples are selection of protein pairs from separate cellular compartments (22) and random selection of protein pairs (23–25). Either of the two strategies has its own limitation. Two proteins localizing to different cellular components could interact with each other (e.g. in the nucleus and cytoplasm, respectively). The negative examples selected by random scheme can be often

*To whom correspondence should be addressed. Tel: +86 10 58805045; Fax: +86 10 58807721; Email: linkui@bnu.edu.cn contaminated with positive ones because of the incomplete protein-interaction network.

To date, protein–protein interaction data do not provide explicit information about the specific regions of the proteins involved in binding or docking. These specific regions, in general only a subset of residues or very short and specific sequence segments (often 3–8 residues) within both interacting proteins, are critical for the highly specific recognition at the contact interface (referred to as the interaction or binding sites) (26–28). Such binding sites are implicated in many fundamental biological processes, including phosphorylation, modification and disease pathways, especially in signaling networks (29–31). Therefore, accurate identification of such interaction sites is essential to understand protein function, and helpful to design and rationalize protein engineering, folding experiments (32–34). Many highly efficient computational methods have been developed to assist the discovery of potential binding sites, especially through mining those protein-interaction datasets produced by high-throughput techniques on a genome-wide scale. In the past few years, most efforts for the prediction of interaction-site pairs were concentrated on finding interaction correlations between domain pairs by statistical analyses (35–43). Nonetheless, it is well known that the actual interaction sites directly responsible for protein binding are probably smaller than the whole domains, and are just subregions of the interacting domains. Recently, several studies have used protein–protein interactions in conjunction with prior biological knowledge to yield a set of putative interacting motif pairs. Li and Li used protein–protein interactions and protein complexes derived from Protein Data Bank (PDB) to identify stable and significant binding motif pairs that have unexpected frequency compared to random in protein-interaction datasets (44). Later, Li *et al.* mined all-versus-all interaction subnetworks to discover motif pairs at interaction sites on a proteome-wide scale (45). Tan *et al.* proposed a novel algorithm, D-MOTIF, to infer correlated motifs from interaction data (46). Yu *et al.* applied the AdaBoost algorithm to predict motif pairs from known interactions and putative non-interacting protein pairs (47). Wang *et al.* proposed a modified model inspired by Deng *et al.* (36) and Riley *et al.* (37) to predict interacting motif/domain pairs, and in particular, the specific binding regions involved in a certain protein interaction (43).

In this study, we focused on identifying motif pairs at interaction sites expected to mediate protein–protein interactions by mining both gold standard positives (GSPs) and gold standard negatives (GSNs) in yeast. Because protein-interaction sites are more conserved than the rest of the protein surface (48), we used short linear peptide motifs to represent the interaction sites (often 3–8 residues) where protein interactions take place. The linear motifs conform to particular sequence patterns indicative of a particular function. Currently, there are several motif databases such as the Eukaryotic Linear Motif (ELM) database (49), PROSITE (50), ScanSite (51) and Minimotif Miner (MnM) (52). Of these, MnM is a newly published motif database with a broad functional spectrum, and its contents were complied from searching the literature or exploring other public databases including PROSITE, ELM and Peptide Cutter. All motifs in MnM have been published and validated with experimental evidence. Because of its high quality, the motifs in MnM were used to annotate the yeast proteins in our study.

The GSP dataset was generated by measuring the relationship strength (including the functional association or the localization proximity) between two different proteins using a relative specificity similarity method. This was achieved by exploring the information buried in the Gene Ontology (GO) and GO annotations in our previous study (8,9). The reconstructed yeast protein–protein interaction map was proved to have a high confidence level when validated using the widely used evaluation dataset compiled from MIPS (53). Four negative datasets were generated by different methods, including a dataset of randomly selected protein pairs, a dataset of protein pairs with different cellular sublocalizations, and two datasets generated with different confidence levels based on the RSS method designed in (9). Furthermore, the quality of the four negative datasets was evaluated and compared. Of these, the one with the best performance was considered as the GSN dataset. To identify putative interacting motif pairs that are statistically overrepresented in their occurrence in the GSPs compared to the GSNs, two distinct statistical tests, the exact binomial test and Fisher's exact test, were integrated. The performance of the predicted results was validated by mapping the inferred motif pairs to three widely used datasets including iPfam (54), DOMINO (55) and the Yeast Core subset in DIP (56). Moreover, we also compared our method with the previously developed methods, and found our method outperformed the others in terms of prediction precision and converge. These results demonstrate that, by incorporating a high-quality negative dataset, our method presents good capability in identifying the interacting motif pairs mediating protein–protein interactions.

MATERIALS AND METHODS

Motif assignments

The motif definitions were drawn from the MnM database. The MnM motif database (the release of Jun 13, 2007) compiles 611 distinct motifs involved in a broad range, such as posttranslational modifications; binding to proteins, nucleic acids or small molecules; protein trafficking; and so on. Information on the subcellular localizations of a motif is also provided, and was utilized as a criterion to filter the false positive motif assignments in this study. We simply specified that if a motif and a protein localize to different subcellular compartments, the motif assignments to the protein be abandoned. We note that the proteins without motif assignments were also discarded. The filtering process is described as follows. First, both the proteins observed in the GSPs and GSNs and the motifs in MnM were annotated with one or more GO cellular compartment (CC) terms. Only if there was a path between one CC term of a protein and one CC term of a motif, was the motif assigned to the protein.

Positive- and negative-interaction datasets

In our previous work (9), we reconstructed a map of potential protein–protein interactions by fully exploring the information contained in the Biological Process (BP) and CC annotations of GO for the yeast genome. The premises of our method were: (a) interacting proteins often function in the same biological process and (b) interacting proteins should exist in close proximity. This was achieved by comparing the relative specificity similarity (RSS) of pairs of GO terms assigned to the two proteins within a GO DAG. The RSS is a new metric of semantic similarity used to score the degree of the functional association or localization proximity between two different proteins. The RSS values for CC and BP ontologies are denoted as RSS^{CC} and RSS^{BP}. We created a GSP dataset using protein pairs with values of RSS^{CC} >0.80 and RSS^{BP} >0.80 based on a new release of GO (the March 2006 release) and the GO annotations derived from SGD (submitted on March 31, 2006), which is now stored in the SPIDer database (8). To improve the quality of the GSP dataset, here we used the more stringent criterion of RSS^{CC} >0.85 and RSS^{BP} >0.85 (referred to as WGSPs). After motif assignments using a cellular compartment filter (as described earlier), the WGSP dataset consisted of 46 031 interacting protein pairs encompassing 2678 proteins. To assess how likely a protein pair in the WGSP dataset was to physically interact with each other, we created a high-quality validation dataset, called 'valid experimental interactions' (VEIs). VEIs combine the binary interactions from the MIPS complexes, the MIPS small-scale physical interactions, and the integrated interactions from de Lichtenberg *et al.* (57). There were 12 345 unique binary interactions among 1905 proteins in the VEIs. The MIPS complexes and the MIPS physical interactions are often used as or as part of the 'gold standard positives' to validate various prediction methods (19,58,59) and are also used to assess high-throughput interaction datasets (60,61). As a result, the WGSP dataset covered about 81% of the VEIs, proving that WGSPs had a high-confidence level. Thereafter, we simply used GSPs to refer to this new GSP dataset (WGSPs).

Four negative datasets assembled by different strategies were constructed in this study. (i) RGSNs: random pairs of proteins that are not known to interact. (ii) SGSNs: as described in (19), the protein pairs in SGSNs were selected from lists of proteins in separate subcellular compartments (cytoplasm, nucleus, mitochondrion and exocytic network) (62) according to the yeast localization data in GO (the details of the construction of SGSNs are available in the Supplementary Materials). According to the distribution of RSS values in the CC and BP ontologies, the RSS^{CC} and RSS^{BP} values were roughly divided into three confidence levels, high (H), medium (M) and low (L) confidence (see Supplementary Materials Figures S1 and S2). Then the other two negative datasets, W1GSNs and W2GSNs, were created based on different combinations of RSS^{CC} and RSS^{BP}. (iii) W1GSNs: protein pairs that have both RSS^{CC} and RSS^{BP} values with low confidence levels, namely the ones localizing in different cellular components and involved in weakly related or unrelated biological processes. (iv) W2GSNs: protein pairs that have RSS^{BP} values with low confidence level and RSS^{CC} values with median or low confidence level. In contrast to W1GSNs, W2GSNs had a larger size by including protein pairs localizing in relatively close cellular components (RSS^{CC} value with median confidence) but involved in weakly related or unrelated biological processes (RSS^{BP} value with low confidence level). Because the number of randomly selected protein pairs is very large, the size of RGSNs was simply chosen to be equal to that of W2GSNs. After motif assignments using the cellular compartment filter, W1GSNs, W2GSNs, RGSNs and SGSNs remained 66 183, 596 669, 645 009 and 3 815 110 protein pairs, respectively. For fair comparison, the four negative datasets were created from the same protein set that comprised 3654 proteins.

Statistical analysis

To measure the overrepresentation of the occurrence of motif pairs in positives compared to negatives, two distinct statistical models for counting the occurrence of motif pairs were adopted. Furthermore, the problem of multiple testing was taken into account in the process of statistical analysis.

One-tailed exact binomial test. The exact binomial test uses the binomial distribution model to compare the rate of the observed occurrence of a motif pair to the expected rate. The motif pairs both significantly overrepresented in the GSPs and significantly underrepresented in the GSNs were determined to be putative interacting motif pairs. Thus, using the *R* statistics package, for a given motif pair two *P-values* were calculated, one corresponding to the statistical significance in the GSPs and the other in the GSNs. Three basic parameters are required for the exact binomial test: the number of successes, the number of trials and the hypothesized probability of success. For a motif pair M_{ij} in protein pair dataset I of size N encompassing n proteins, the three parameters respectively correspond to X_{ij} (the observed number of protein pairs containing M_{ij}, where one protein contains the motif i and its partner contains the other motif j), N (the size of the protein pair dataset I) and Ef_{ij} (the expected frequency of protein pairs containing M_{ij}). Ef_{ij} was calculated as S_{ij}/C_n^2, where C_n^2 *is* the size of the universe of protein pairs collected from the n proteins (homo-pairs were excluded) and S_{ij} is the number of all the protein pairs containing M_{ij} in the universe. The exact binomial test is performed to evaluate significant differences in the rate of the occurrence of motif pairs, and thus is particularly good at detecting increased prevalence of common motif pairs.

One-tailed Fisher's exact test. In contrast to the exact binomial test, the Fisher's exact test uses a hypergeometric distribution model to compare the proportion of protein pairs containing a motif pair in the GSPs to that in the GSNs, and therefore is good at detecting rare motif pairs that occur less frequently in interacting protein pairs. For the Fisher's exact test, a 2 × 2 contingency table of frequency is created for each motif pair, in which the two rows represent the GSPs and GSNs, respectively, and the two columns represent the numbers of protein pairs containing the given motif pair and the ones not containing the motif pair, respectively. Using the *R* statistics package, each motif pair is assigned with a *P-value*.

Multiple testing problem. The *q-value* method proposed by Storey (63,64) was employed to control the false discovery rate (FDR). The *q-value* measures the expected proportion of false positives incurred when a test is called significant. Similar to a *P-value*, the *q-value* can be considered a measure of statistical significance. We used QVALUE software, which takes a list of *P-values* resulting from the simultaneous tests as input and estimates their *q-values* (63). The *q-value* can be calculated for each test and ranked in ascending order. In practice, a cutoff for null hypothesis rejection was set to 0.05 to ensure a 5% FDR.

Validation datasets

Currently, comprehensive interacting motif pair data do not yet exist and are difficult to assemble. Fortunately, there are several high-quality databases of interaction sites, such as iPfam, DOMINO and the *Saccharomyces cerevisiae* core subset in DIP (Yeast Core). Here, we defined a pair of sequence segments with exact start and end positions to represent an interaction-site pair. iPfam is a popular database of domain–domain interactions derived from the protein complexes in PDB (54). It contains 3020 domain–domain interactions (version 20). DOMINO is a database of domain–peptide interactions storing more than 3900 annotated interactions with experimentally verified evidence (55), from which only the segment pairs with both exactly annotated start and end positions were used in this study. In addition, a high-confidence Yeast Core dataset of protein interactions in DIP generated by merging several high-quality subsets from experimental and computational validation (56) was used. The sequences of proteins composing the interacting protein pairs could be regarded as the maximal potential interaction regions. The core dataset (the release of 7 January 2007) contains 17 420 protein–protein interactions encompassing 4909 proteins. Note that for iPfam and DOMINO, only the segment pairs in *S. cerevisiae* were chosen.

We defined that a motif can be mapped to a sequence segment if one instance of the motif is nested by the segment. Then we defined that a motif pair can be mapped to a segment pair if both members of the motif pair can be mapped to those of the segment pair. Finally, after motif assignment using the cellular compartment filter, the respective numbers of segment pairs for the iPfam, DOMINO and Yeast Core datasets were 351, 392 and 12 680, respectively.

The validation of the inferred motif pairs was performed by estimating their positive predictive values (PPVs) and sensitivities (SNs). The PPV was calculated as TP/(TP + FP), where true positives (TP) and false positives (FP) were estimated with respect to each validation dataset. As negative datasets of motif pairs do not yet exist, we simply defined PPV as the proportion of the inferred motif pairs overlapping with each validation dataset. The SN, calculated as TP/P (P being the size of the validation dataset), was simply defined as the proportion of the segment pairs in each validation dataset overlapping with the inferred motif pairs.

Randomizing simulation

Obviously, a good prediction system should contain more inferred motif pairs mapped to the validation datasets than expected at random. For evaluating the enrichment of our inferred motif pairs in the validation datasets, we attached a measure of statistical significance to the overlaps. As the distribution of the overlaps is unknown, we estimated the significance by randomizing the simulation process. To do so, for each validation dataset, we randomly generated 1000 datasets of segment pairs collected from the segments composing the validation dataset. The size of each randomly generated dataset is the same as that of the validation dataset. Both the PPVs and SNs of the inferred motif pairs with the validation datasets were assigned empirical *P-values*. The empirical *P-value* was calculated as the proportion of the simulated datasets with an equal or larger PPV (or SN) than the observed one.

RESULTS

Assessment of four negative datasets

To evaluate the effect of the four negative datasets (RGSNs, SGSNs, W1GSNs and W2GSNs) on identifying interacting motif pairs, we compared their respective inferred motif pairs with interaction-site pairs derived from the three reference databases, iPfam, DOMINO and the Yeast Core in DIP. We used the exact binomial test to predict the putative interacting motif pairs mining from each negative dataset. It should be noted that as the interaction sites in Yeast Core were roughly defined as the whole protein sequences, iPfam and DOMINO have more accurate definitions of interaction sites than Yeast Core. Therefore, the evaluation of the different negative datasets (and the assessments thereafter) depended mainly on the validation results derived from iPfam and DOMINO, while the validation results from DIP can be considered as auxiliary evidence.

The respective numbers of the inferred motif pairs mining from RGSNs, SGSNs, W1GSNs and W2GSNs were 38, 4593, 3684 and 1762. Tables 1 and 2 list validation result statistics of the four negative datasets. Surprisingly, only a small number of motif pairs were predicted by RGSNs, much fewer than from the other negative datasets. Although the PPVs of RGSNs were highest (Table 1), the SNs were much lower than those of the other datasets (Table 2). Moreover, its SN with DOMINO and PPVs with iPfam and Yeast Core were not significant. The reason may be that, because of a lack of

Table 1. Positive predictive values (PPVs) of the motif pairs inferred by the exact binomial test from the four negative datasets

	W1GSNs		SGSNs		W2GSNs		RGSNs	
	PPV (%)	*P*-value	PPV (%)	*P*-value	PPV (%)	*P*-value	PPV (%)	*P*-value
DOMINO	15.61	0.006	13.52	0.013	9.36	0.018	78.95	0.025
iPfam	14.69	0.032	12.61	0.014	11.80	0.012	81.58	0.957
Yeast Core	98.24	0.930	95.17	0.085	96.54	0.059	100.00	1.000

PPV was calculated as TP/(TP + FP), where true positives (TP) and false positives (FP) were estimated with respect to each validation dataset. Here, PPV was defined as the proportion of the inferred motif pairs overlapping with each validation dataset.

Table 2. Sensitivities (SNs) of the motif pairs inferred by the exact binomial test from the four negative datasets

	W1GSNs		SGSNs		W2GSNs		RGSNs	
	SN (%)	*P*-value	SN (%)	*P*-value	SN (%)	*P*-value	SN (%)	*P*-value
DOMINO	75.77	<0.001	75.77	<0.001	5.87	1.000	1.79	0.750
iPfam	95.16	<0.001	95.16	<0.001	60.40	<0.001	45.30	<0.001
Yeast Core	97.93	<0.001	98.23	<0.001	87.82	<0.001	53.24	<0.001

The SN, calculated as TP/P (P being the size of validation dataset), was simply defined as the proportion of the segment pairs in each validation dataset overlapping with the inferred motif pairs.

biological significance, compared with other methods the random selection method would be more likely to choose positive examples or protein pairs with similar attributes as positives (e.g. with close proximity or related biological process).

Figure 1 shows a comparison of the PPVs and SNs (defined in 'Materials and Methods' section) of SGSNs, W1GSNs and W2GSNs. We observed that W1GSNs generally outperformed SGSNs and W2GSNs both in terms of PPVs and SNs, and SGSNs came second. The superior performance of W1GSNs to SGSNs is due to W1GSNs' stricter generation criteria (involved in both different biological processes and different cellular compartments) than for SGSNs (only considering different cellular compartments). Compared with W1GSNs (or SGSNs), the lower performance of W2GSNs may be due to the inclusion of the protein pairs with median-confidence RSS^{CC}, implying the main effect of the localization proximity on the capability of negative datasets in identifying interacting motif pairs. Comparison of the four negative datasets based on the motif pairs inferred by the Fisher's exact test was also performed. We obtained the similar results that generally W1GSNs performed the best (see Supplementary Materials Tables S2a and S2b, Figure S3). Finally, considering that W1GSNs were generated using the most stringent criteria and produced in the same system as WGSPs, we chose W1GSNs as the GSNs for predicting interacting motif pairs. Thereafter, we simply used GSNs to refer to W1GSNs.

In addition, we found that for the four negative datasets, the PPV values of Yeast Core were not significant. A plausible explanation may be that as the definition of interaction sites of Yeast Core is rather general (the whole protein sequence), while linear motifs are short and less specific in contrast to domains, it would lead to frequent nonfunctional (or random) motif assignments along proteins; consequently a number of motif pairs may appear randomly in the simulation datasets of Yeast Core, which would make the validation results of Yeast Core non-significant.

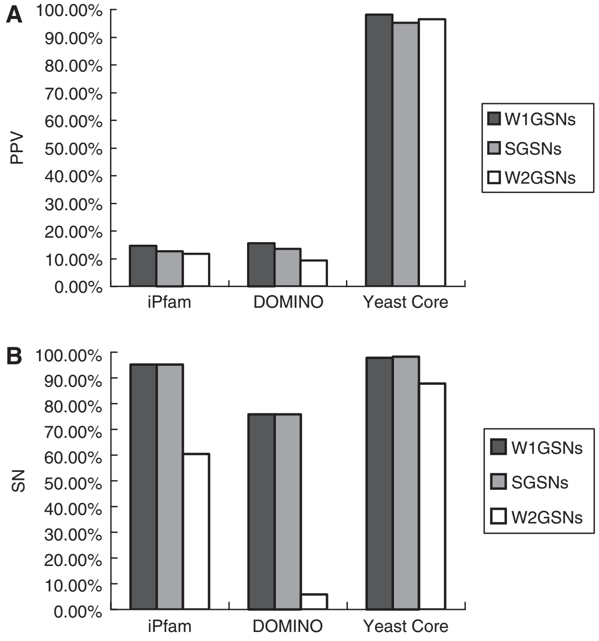

Figure 1. Comparison of the (**A**) positive predictive values [PPVs, defined as TP/(TP + FP)] and (**B**) sensitivities (SNs) of the three negative datasets W1GSNs, W2GSNs and SGSNs. W1GSNs and W2GSNs were generated using our previously described RSS method (9). The RSS is a new metric of semantic similarity used to score the degree of the functional association or localization proximity between two different proteins. W1GSNs comprised protein pairs with low-confidence RSS^{BP} and low-confidence RSS^{CC}; W2GSNs comprised the protein pairs with low-confidence RSS^{BP} and low- or median-confidence RSS^{CC} values. SGSNs were generated using the method of selecting protein pairs with different subcellular localizations.

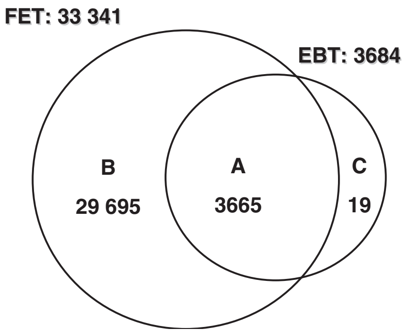

Figure 2. A Venn diagram of the numbers of motif pairs inferred by the exact binomial test and the Fisher's exact test. These motif pairs can be divided into three data groups: (**A**) the intersection between the dataset inferred by the exact binomial test (EBT) and the dataset inferred by the Fisher's exact test (FET); (**B**) the portion inferred only by the Fisher's exact test and (**C**) the portion inferred only by the exact binomial test.

Table 3. Statistical analysis of the validation results of the motif pairs inferred by the Fisher's exact test

	PPV (%)	*P*-value[a]	SN (%)	*P*-value[b]
DOMINO	16.13	0.027	89.54	<0.001
iPfam	31.90	0.190	99.15	<0.001
Yeast Core	98.41	0.506	99.95	<0.001

[a]The empirical *P*-value for the PPVs with the validation datasets.
[b]The empirical *P*-value for the SNs with the validation datasets.

Inference of putative protein interacting motif pairs

We implemented both the exact binomial test and the Fisher's exact test to assess the statistical significance of the overrepresentation of co-occurring motif pairs in the GSPs compared to the GSNs. For the exact binomial test with *q-value* <0.01, 3684 putative interacting motif pairs both significantly overrepresented in the GSPs and underrepresented in the GSNs (referred to as the EBT dataset) were detected. For the Fisher's exact test with *q-value* <0.01, 33 341 putative interacting motif pairs (denoted as the FET dataset) were obtained. And 3665 motif pairs overlapped between EBT and FET, whereas, 29 695 were inferred solely by the Fisher's exact test and 19 were inferred solely by the exact binomial test. Thereafter, these three groups of inferred motif pairs are denoted as A, B and C, respectively (Figure 2).

We found that FET contained a much larger number of motif pairs than EBT. The reason is that the Fisher's exact test can detect both common and rare motif pairs and therefore be more sensitive (higher SN) than the exact binomial test as shown in Tables 2 and 3. Although the PPVs of FET were higher than EBT, the empirical *P-values* for the PPVs with iPfam and DOMINO were less than those of EBT, especially for iPfam, the result was not significant (Tables 1 and 3). These results indicate that the higher sensitive of FET may be accompanied with higher false positives. As shown in Tables 1 and 3, we noted that the PPVs for iPfam and DOMINO were at a low level compared with the Yeast Core dataset. A plausible explanation may be that these two datasets are relatively incomplete, for the domain–domain interactions in iPfam are observed in the protein complexes with known 3D structures, and DOMINO only collects interactions with experimentally verified evidence. As expected, only a small fraction of biologically occurring interaction-site pairs was sampled.

Assembly of an interacting motif pair dataset with high confidence

As the exact binomial test does well in detecting common motif pairs, and the Fisher's exact test is effective for detecting rare motif pairs, the two interacting motif pair datasets inferred by the distinct statistical methods were combined. Before doing this, we defined that a motif pair can be assigned with one of the three evidence types corresponding to the three validation databases, iPfam, DOMINO and Yeast Core, if it can be mapped to one of the validation datasets. According to the number of evidence types, motif pairs can be divided into four groups: no evidence (evi0), exactly only one evidence type (evi1), two evidence types (evi2) and three evidence types (evi3). Intuitively, the larger the number of evidence types a motif pair has, the greater the confidence level for the motif pair.

First, as 99.48% (3665 out of 3684) of the motif pairs in EBT were also predicted by FET method, we assembled EBT into the final interacting motif pair dataset. To increase the coverage of the interacting motif pair dataset, the motif pairs solely inferred from the Fisher's exact test (group B) were also considered as a candidate set (Figure 2). Because of the propensity of FET to contain more false positives as described earlier, we were interested in those motif pairs that appear underrepresented with significance in the GSPs and overrepresented in the GSNs (but without significance), where the statistical significance was according to the *q-values* derived from the exact binomial test. We defined that a motif pair M_{ij} is 'overrepresented in the GSPs but without significance' if $Nobs_{ij} > Nexp_{ij}$ and *q-value* ≥0.01, where $Nobs_{ij}$ is the observed number of protein pairs containing M_{ij} and $Nexp_{ij}$ is the expected number of protein pairs containing M_{ij}. $Nexp_{ij}$ is calculated as $Ef_{ij} \times N$, where Ef_{ij} *is* the expected frequency of protein pairs containing M_{ij} (defined in the 'Materials and Methods' section) and N is the size of protein pair dataset. As a result, 5731 motif pairs (denoted as filtered group B) were extracted and assembled into the final interacting motif pair dataset.

Therefore, two groups of motif pairs, EBT and filtered group B, were incorporated into the set of high-confidence interacting motif pairs [denoted as Interacting Motif Pairs (IMP)]. IMP contained 9415 motif pairs in total. We found that only 2.25% (212/9415) of motif pairs in IMP had no evidence (Figure 3), and IMP covered 96.01, 78.57 and 98.51% of iPfam, DOMINO and Yeast Core, respectively.

Ranking the inferred motif pairs with high confidence

We ranked the motif pairs according to the *q*-values from the exact binomial test in GSPs and the *q*-values

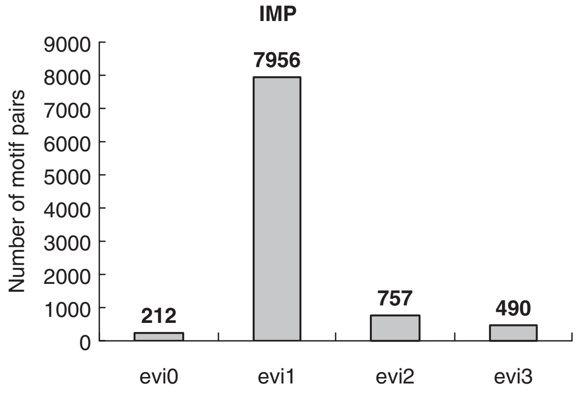

Figure 3. The distribution of the inferred motif pairs (IMPs) with different numbers of evidence types.

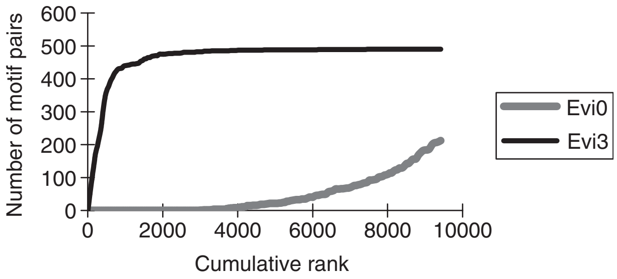

Figure 4. The distribution of the inferred motif pairs in different cumulative ranks. The number of the inferred motif pairs is plotted against their cumulative ranks. At the various rank cutoffs, the numbers of the motif pairs with no evidence (evi0) and with three evidence types (evi3) were counted, respectively.

from the Fisher's exact test in ascending order, respectively. Then we compared the performance of the two ranking methods. The *q*-value of the Fisher's exact test had the better performance and was used as our ranking scheme (see Supplementary Materials Figure S4). The reason may be that some specific and rare motif pairs that are likely to be true interacting motif pairs would be assigned a higher rank in the Fisher's exact test. We expected that motif pairs with more evidence types, which are more likely to interact with each other, should rank higher. This can be determined by analyzing the distribution of the frequency of the motif pairs in IMP with different evidence types. As shown in Figure 4, at the various rank cutoffs, the numbers of motif pairs within evi0 and evi3 were counted. Intuitively, the motif pairs within evi0 and evi3 can be regarded as the least and the most reliable, respectively. The motif pairs within evi0 appear seldom among the top ranks, of which the highest rank is 2972; while the majority of the motif pairs within evi3 are top-ranked (i.e. about 50% are among the top 216).

Interesting motif pairs inferred with no validation evidence

Because of the incompleteness of the validation datasets, motif pairs inferred without evidence might truly bind to each other to mediate protein–protein interactions. In order to find whether there is evidence to indirectly support their binding function, we mapped the 212 motif pairs without evidence in IMP to the datasets of interaction-site pairs in the other species. To this end, three confidence datasets were adopted, iPfam, DOMINO and the dataset compiled from Pawson lab (http://pawsonlab.mshri.on.ca/). As a result, 93 motif pairs have been mapped to the interaction-site pairs of the other species. For instance, the motif pair composed of the motif *FGRA* [MnM:PBMDNA00004A] and the motif *Y-D/K/N-H/F/R-P/V/L* [MnM:PBMSH200020B], occurs in the physical interacting regions of a DNA-binding protein in *Methanopyrus kandleri* [PDBID:1f1e] (Figure 5). Another instance is that the motif *V/Y/F-?-I/V/A>* [MnM:PBMPDZ00002A], a PDZ Class II binding motif, was predicted to interact with the Motif *CPV* [MnM:PBMMHL00001A] occurring in a PDZ domain in *Mus musculus* (65).

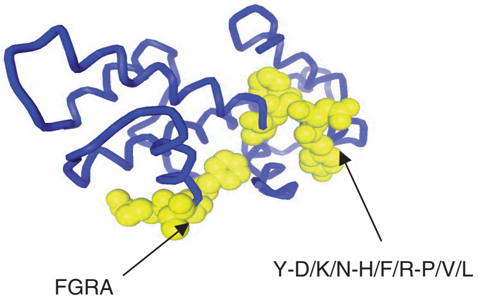

Figure 5. Three-dimensional structure of a motif pair without evidence in yeast but found in the physical interacting regions of DNA-binding protein in *Methanopyrus kandleri* [PDB:1f1e].

DISCUSSION

Comparison with previously developed methods

MLE (36), DPEA (37) and a simple association method were compared with our method. The measure of the simple association method is defined as the fraction of the interacting protein pairs among all of the protein pairs containing a given motif pair M_{ij}. The measure of the MLE method is the estimated value of the probability of an interacting motif pair $\Pr(M_{ij} = 1)$ by using the expectation maximization algorithm (EM) to maximize the expectation of observing a given protein interaction network. We calculated it as Deng *et al.* (36). We also used an extended measure of MLE provided by Lee *et al.* (41), the expected number of occurrences of motif pairs. It is defined as $N_{ij} \times \Pr(M_{ij} = 1)$ where N_{ij} is the number of all protein pairs containing a motif pair M_{ij}. The DPEA method is based on computing an *E-value*, which measures how disallowing the given motif pair reduces the likelihood of a protein–protein interaction network (37). The four measures are referred to as Frequency, Probability, Expectation and *E*-value, respectively. The measure of our method is referred to as Qvalue. The power of the different methods was evaluated by plotting the curves of their PPV values versus the top percent rank in the validation datasets iPfam and DOMINO (Figure 6). We observed

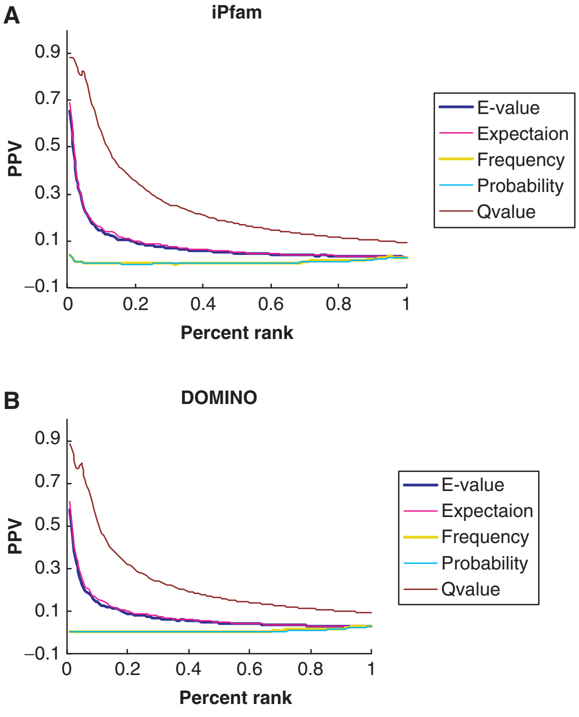

Figure 6. The relationship between the top percent rank versus the positive predictive value (PPV) estimated by iPfam (**A**) and DOMINO (**B**). Five measures of prediction methods were assessed. 'Frequency' is the measure of a simple association method that scores the fraction of the interacting protein pairs among all of the protein pairs containing a given motif pair. 'Probability' is the measure of the MLE method to score the probability of an interacting motif pair (36). 'Expectation' is an extended measure of the MLE provided by Lee *et al.* (41) that scores the expected number of occurrences of motif pairs. '*E*-value' is the measure of the DPEA method that measures how disallowing the given motif pair reduces the likelihood of a protein–protein interaction network (37). The measure of our method is referred to as the 'Qvalue', which is calculated using the Fisher's exact test.

that the Qvalue outperformed the other methods in the two validation datasets. The *E*-value and Expectation had similar performance and came second, and the Probability and Frequency performed the worst, which were also observed in (37,40,41). A plot similar to Figure 6, depicting the relationship between SN versus the top percent rank is available as Supplementary Materials Figure S5. Similar results were obtained.

Our Qvalue method is an association method. The dominance of the Qvalue method over the others could be attributed to two main reasons. First, compared with the simple association method, Frequency, the Qvalue method uses more stringent statistical tests to find motif pairs with significant occurrence. Second, an advantage of the complicated methods (MLE and DPEA) is that they take into account the mutual impact of multiple motif pairs coexisting in an interacting protein pair on the interaction of the protein pair. However, in contrast to domains, motif assignments may introduce much more noise because of the lower specificity of linear motifs, so the advantage of the complicated methods in considering the mutual effect of multiple motif pairs may be impaired. Moreover, the complicated methods have so many parameters to be tuned that they are more likely to be affected by this noise. In such a situation, a simple model with strict statistical analysis may be more suitable. However, we should note that because we have not compared our method with these complicated methods in predicting domain–domain interaction, the results can only suggest that our method performs better than the complicated methods when identifying interacting motif pairs.

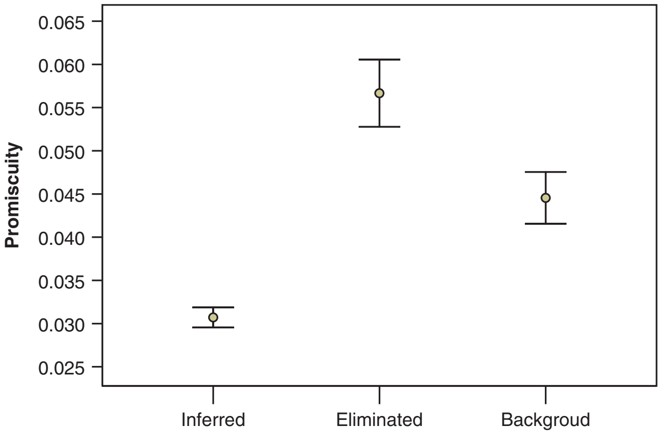

Figure 7. Comparison of the promiscuity of the motif pairs among the three datasets: 'Inferred'—the motif pairs were both significantly overrepresented in the GSPs and underrepresented in the GSNs, 'Eliminated'—the motif pairs were significantly overrepresented in the GSPs but did not satisfy the criterion of 'significantly underrepresented in the GSNs' and 'Background'—the motif pairs were significantly overrepresented in the GSPs. The promiscuity of a motif pair was measured by $\#Pairs_{observed}/\#Pairs_{possible}$, where $\#Pairs_{observed}$ is the number of the observed interacting protein pairs containing the motif pair, and $\#Pairs_{possible}$ is the number of all the possible protein pairs containing the motif pair.

The effectiveness of a high-quality negative dataset on inference performance

We took the exact binomial test as an example to investigate the effectiveness of a high-quality negative dataset. A serious problem underlying methods of inferring interacting motif pairs is that promiscuous motif pairs are scored highly because of the frequency of their occurrence, but not to because of the specific topology of the network (37). We wondered, by using a high-quality negative dataset, whether the overprediction of promiscuous interactions could be controlled. This is based on the assumption that through incorporating high-quality negatives, some false positives could be reduced by eliminating the motif pairs significantly overrepresented in both the GSPs and GSNs, and that these eliminated motif pairs usually occur promiscuously in many if not most interacting proteins. In total, there were 5101 motif pairs overrepresented in the GSPs regardless of their occurrences in the GSNs (called 'Background'), and 1417 (about 27%) were eliminated by the GSNs (called 'Eliminated'). To this end, we tested this assumption by comparing the promiscuity of motif pairs among the three datasets, 'Background', 'Inferred' (EBT) and 'Eliminated'. As shown in Figure 7, the promiscuity of 'Inferred' was significantly less than that of 'Eliminated'

(Mann–Whitney *U*-test, *P-value* < 2.2e-16) and that of 'Background' (*P-value* < 2.2e-16), and the promiscuity of 'Eliminated' was significantly higher than that of 'Background' (*P-value* < 2.2e-16), suggesting the data eliminated by the GSNs contain the most promiscuous interactions. In addition, we also mapped these eliminated motif pairs to the validation datasets (see Supplementary Materials Table S3). We found that the PPVs and SNs of these eliminated motif pairs were much less than PPVs and SNs of those mining from both the GSPs and GSNs (EBT dataset, Tables 1 and 2), indicating these eliminated motif pairs may contain high false positives (for details see Supplementary Materials). These results suggest that a high-quality negative dataset has a large effect on decreasing motif pairs with promiscuous interactions, and plays a critical role in the inference of interacting motif pairs with high confidence.

Caveats on our method

There are several underlying limitations in our approach. (i) In contrast to domains, linear motifs are difficult to detect experimentally or computationally because of their short length and some degree of degeneracy. Therefore, existing motif databases are far from comprehensive, and thus the use of these predefined patterns will reduce the motif search space to enable motif pair mining in large interaction networks. (ii) Another problem is non-functional false positive assignments, which is a serious consideration in motif assignments. In this study, we used information regarding subcellular components to filter out putative false positive assignments, but the effectiveness of such a strategy may be still limited. We expect to integrate other information such as species information and evolutionary conservation to reduce false positive rates in our future work. (iii) As our work was only based on *S. cerevisiae*, some motif pairs specific to other species or those appearing rarely in yeast could not be detected by our method. Thus, in the future, our interacting motif pair mining method will be extended to other organisms, and thus both the accuracy and coverage of our prediction system should be improved greatly.

Finally, we should note that the statistical significance used in our method is not equivalent to biological function. Not every protein with one motif of our inferred interacting motif pair is expected to interact with another protein with the other motif of the pair. The inferred motif pairs may indirectly mediate protein interactions, or help shape the structure of proteins. In any case, the motif pairs predicted by our method can be used to direct new experimental interaction screens, in both yeast and other species, through which the search space of putative interacting protein pairs would be greatly reduced.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for their invaluable comments. This research was supported by NSFC (Grants 30770445 and 30571037) and by Beijing Normal University. Funding to pay the Open Access publication charges for this article was provided by Beijing Normal University.

Conflict of interest statement. None declared.

REFERENCES

1. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan C. elegans. *Science*, **303**, 540–543.
2. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.
3. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The protein-protein interaction map of Helicobacter pylori. *Nature*, **409**, 211–215.
4. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
5. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
6. Xia,Y., Yu,H., Jansen,R., Seringhaus,M., Baxter,S., Greenbaum,D., Zhao,H. and Gerstein,M. (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.*, **73**, 1051–1087.
7. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 368–373.
8. Wu,X., Zhu,L., Guo,J., Fu,C., Zhou,H., Dong,D., Li,Z., Zhang,D.Y. and Lin,K. (2006) SPIDer: Saccharomyces protein-protein interaction database. *BMC Bioinformatics*, **7(Suppl 5)**, S16.
9. Wu,X., Zhu,L., Guo,J., Zhang,D.Y. and Lin,K. (2006) Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. *Nucleic Acids Res.*, **34**, 2137–2150.
10. Bader,G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND–the biomolecular interaction network database. *Nucleic Acids Res.*, **29**, 242–245.
11. Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.
12. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res.*, **32**, D452–D455.
13. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
14. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
15. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
16. Jeong,H., Mason,S.P., Barabasi,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
17. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. *Nat Biotechnol.*, **21**, 697–700.
18. Kim,W.K., Henschel,A., Winter,C. and Schroeder,M. (2006) The many faces of protein-protein interactions: a compendium of interface geometry. *PLoS Comput. Biol.*, **2**, e124.
19. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
20. Sprinzak,E., Sattath,S. and Margalit,H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.*, **327**, 919–923.

21. Hart,G.T., Ramani,A. and Marcotte,E. (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120.
22. Jansen,R. and Gerstein,M. (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.*, **7**, 535–545.
23. Chen,X.W. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.
24. Shen,J., Zhang,J., Luo,X., Zhu,W., Yu,K., Chen,K., Li,Y. and Jiang,H. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.
25. Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein-protein interactions. *Bioinformatics*, **21(Suppl 1)**, i38–i46.
26. Forrest,J.C., Campbell,J.A., Schelling,P., Stehle,T. and Dermody,T.S. (2003) Structure-function analysis of reovirus binding to junctional adhesion molecule 1. Implications for the mechanism of reovirus attachment. *J. Biol. Chem.*, **278**, 48434–48444.
27. Zhang,Y., Rassa,J.C., deObaldia,M.E., Albritton,L.M. and Ross,S.R. (2003) Identification of the receptor binding domain of the mouse mammary tumor virus envelope protein. *J. Virol.*, **77**, 10468–10478.
28. Kim,W.K., Henschel,A., Winter,C. and Schroeder,M. (2006) The many faces of protein-protein interactions: a compendium of interface geometry. *Plos Comput. Biol.*, **2**, 1151–1164.
29. Mayer,B.J. (2001) SH3 domains: complexity in moderation. *J. Cell Sci.*, **114**, 1253–1263.
30. Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett.*, **579**, 3342–3345.
31. Yaffe,M.B. (2002) Phosphotyrosine-binding domains in signal transduction. *Nat. Rev. Mol. Cell. Biol.*, **3**, 177–186.
32. Lichtarge,O., Sowa,M.E. and Philippi,A. (2002) Evolutionary traces of functional surfaces along G protein signaling pathway. *Methods Enzymol.*, **344**, 536–556.
33. Loregian,A. and Palu,G. (2005) Disruption of protein-protein interactions: towards new targets for chemotherapy. *J. Cell Physiol.*, **204**, 750–762.
34. Arkin,M.R., Randal,M., DeLano,W.L., Hyde,J., Luong,T.N., Oslob,J.D., Raphael,D.R., Taylor,L., Wang,J., McDowell,R.S. *et al.* (2003) Binding of small molecules to an adaptive protein-protein interface. *Proc. Natl Acad. Sci. USA*, **100**, 1603–1608.
35. Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.
36. Deng,M., Mehta,S., Sun,F. and Chen,T. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.
37. Riley,R., Lee,C., Sabatti,C. and Eisenberg,D. (2005) Inferring protein domain interactions from databases of interacting proteins. *Genome Biol.*, **6**, R89.
38. Nye,T.M., Berzuini,C., Gilks,W.R., Babu,M.M. and Teichmann,S.A. (2005) Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, **21**, 993–1001.
39. Ng,S.-K., Zhang,Z. and Tan,S.-H. (2003) Integrative approach for computationally inferring protein domain interactions. *Bioinformatics*, **19**, 923–929.
40. Guimaraes,K., Jothi,R., Zotenko,E. and Przytycka,T. (2006) Predicting domain-domain interactions using a parsimony approach. *Genome Biol.*, **7**, R104.
41. Lee,H., Deng,M., Sun,F. and Chen,T. (2006) An integrated approach to the prediction of domain-domain interactions. *BMC Bioinformatics*, **7**, 269.
42. Chen,X.-W. and Liu,M. (2005) Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, **21**, 4394–4400.
43. Wang,H., Segal,E., Ben-Hur,A., Li,Q.R., Vidal,M. and Koller,D. (2007) InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biol.*, **8**, R192.
44. Li,H. and Li,J. (2005) Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics*, **21**, 314–324.
45. Li,H., Li,J. and Wong,L. (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, **22**, 989–996.
46. Tan,S.H., Hugo,W., Sung,W.K. and Ng,S.K. (2006) A correlated motif approach for finding short linear motifs from protein interaction networks. *BMC Bioinformatics*, **7**, 502.
47. Yu,H., Qian,M.P. and Deng,M.H. (2006) Using a Stochastic AdaBoost algorithm to discover interactome motif pairs from sequences. *Lecture Notes in Comput. Sci.*, **4115**, 622–630.
48. Caffrey,D.R., Somaroo,S., Hughes,J.D., Mintseris,J. and Huang,E.S. (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*, **13**, 190–202.
49. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M.A., Ausiello,G., Brannetti,B., Costantini,A. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
50. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
51. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
52. Balla,S., Thapar,V., Verma,S., Luong,T., Faghri,T., Huang,C.H., Rajasekaran,S., del Campo,J.J., Shinn,J.H., Mohler,W.A. *et al.* (2006) Minimotif Miner: a tool for investigating protein function. *Nat. Methods*, **3**, 175–177.
53. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al.* (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.
54. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
55. Ceol,A., Chatr-aryamontri,A., Santonico,E., Sacco,R., Castagnoli,L. and Cesareni,G. (2007) DOMINO: a database of domain-peptide interactions. *Nucleic Acids Res.*, **35**, D557–560.
56. Deane,C.M., Salwinski,L., Xenarios,I. and Eisenberg,D. (2002) Protein interactions: two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics*, M100037–MCP100200.
57. de Lichtenberg,U., Jensen,L.J., Brunak,S. and Bork,P. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.
58. Lu,L.J., Xia,Y., Paccanaro,A., Yu,H. and Gerstein,M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.
59. Patil,A. and Nakamura,H. (2005) Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics*, **6**, 100.
60. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
61. Edwards,A.M., Kus,B., Jansen,R., Greenbaum,D., Greenblatt,J. and Gerstein,M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**, 529–536.
62. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.*, **16**, 707–719.
63. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
64. Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B – Stat. Methodol.*, **64**, 479–498.
65. Hamazaki,Y., Itoh,M., Sasaki,H., Furuse,M. and Tsukita,S. (2002) Multi-PDZ domain protein 1 (MUPP1) is concentrated at tight junctions through its possible interaction with claudin-1 and junctional adhesion molecule. *J. Biol. Chem.*, **277**, 455–461.