Published online April 26, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 7 2137–2150
doi:10.1093/nar/gkl219

Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations

Xiaomei Wu, Lei Zhu, Jie Guo, 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 January 14, 2006; Revised March 7, 2006; Accepted March 24, 2006

ABSTRACT

A map of protein–protein interactions provides valuable insight into the cellular function and machinery of a proteome. By measuring the similarity between two Gene Ontology (GO) terms with a relative specificity semantic relation, here, we proposed a new method of reconstructing a yeast protein–protein interaction map that is solely based on the GO annotations. The method was validated using high-quality interaction datasets for its effectiveness. Based on a *Z*-score analysis, a positive dataset and a negative dataset for protein–protein interactions were derived. Moreover, a gold standard positive (GSP) dataset with the highest level of confidence that covered 78% of the high-quality interaction dataset and a gold standard negative (GSN) dataset with the lowest level of confidence were derived. In addition, we assessed four high-throughput experimental interaction datasets using the positives and the negatives as well as GSPs and GSNs. Our predicted network reconstructed from GSPs consists of 40 753 interactions among 2259 proteins, and forms 16 connected components. We mapped all of the MIPS complexes except for homodimers onto the predicted network. As a result, ~35% of complexes were identified interconnected. For seven complexes, we also identified some nonmember proteins that may be functionally related to the complexes concerned. This analysis is expected to provide a new approach for predicting the protein–protein interaction maps from other completely sequenced genomes with high-quality GO-based annotations.

INTRODUCTION

One of the main goals of functional genomics is to determine the function of genes predicted from the completely sequenced genomes. In the past decade, massive amounts of biological data have been accumulated from genome sequencing as well as from transcriptomes, proteomes and interactomes. It is a challenging task to integrate such relevant data sources to represent the comprehensive knowledge of genes within and between genomes, which provide specialized information to describe the biological roles of the products of genes. The Gene Ontology (GO) (1) is one such resource that is becoming the *de facto* standard for facilitating information search tasks across databases and for annotating gene products (2). It has been successfully used in protein classification, such as in *Photobacterium profundum* (3), *Plasmodium falciparum* (4), *Drosophila*, *Anopheles* (5–7), *Oryza sativa* (8), as well as *Pan troglodytes* and *Homo sapiens* (9,10). It can also be used in describing gene expression clustering results to explain why a cluster of genes shares a similar expression pattern (11).

The GO has been developed to offer controlled vocabularies for aiding in the annotation of molecular attributes for different model organisms. Three structured ontologies have been proposed, which allow the description of molecular function (MF), biological process (BP) and cellular component (CC). Each ontology is structured as a directed acyclic graph (DAG), which differs from hierarchies in that a ‘child’ (more specialized term) can have many ‘parents’ (less specialized terms or more general terms) and child terms are instances or components of parent terms. Thus, the information derived from the GO must be useful in developing new predictive systems, which may be integrated with other models in large-scale genomic research. Currently, originating from the GO, several functional association predictors have been constructed, which can be roughly grouped into two categories. The techniques in the first category are used to assess the functional

*To whom correspondence should be addressed. Tel: +86 10 58805045; Fax: +86 10 58807721; Email: linkui@bnu.edu.cn associations between proteins in terms of the shared GO terms in a controlled vocabulary system (12–15). However, they are restricted to protein pairs with the same annotations. Techniques from the second category assess the functional associations between proteins using the semantic similarity measures of pairs of terms assigned to them based on either information content (16) or GO structures (17). These two methods in the second category use very similar definitions for the similarity measure for GO annotations, although they treat the specificity of the most recent common ancestor (MRCA) of two GO terms in different ways (17). Motivated by the two methods in the latter class, in this work, we constructed a new functional predictor to systematically predict the map of potential physical interactions between yeast proteins by fully exploring the knowledge buried in two GO annotations for the yeast genome, namely, the BP and CC annotations. Our method is explicitly based upon Wu's similarity measure for GO annotations (17) and is extended to take the relative specificities of GO annotations into account within a given GO structure (see Materials and Methods). Our premise is straightforward from the following two observations: (i) interacting proteins often function in the same biological process, which assumes that two proteins acting in the same biological process are more likely to interact than two proteins involved in different processes, and moreover, proteins functioning in specific biological processes should be more likely to interact than proteins functioning in general processes (14,18–20); (ii) to interact physically, proteins must exist in close proximity, at least transiently, which suggests that co-localization may serves as an useful predictor for protein interactions (19,21).

Since proteins perform their functions by interacting with one another and with other biomolecules, reconstructing a map of the protein–protein interactions of a cell is an important first step toward understanding protein function and cellular behavior (22,23). Recently, genome-scale protein interaction networks have been experimentally determined for *Caenorhabditis elegans* (24), *Drosophila melanogaster* (25), *Helicobacter pylori* (26), *H.sapiens* (12,14,27), and *Saccharomyces cerevisiae* (28–31). Although these experimental techniques have drastically improved our knowledge of protein interactions, the datasets generated from these studies are often noisy and incomplete (32,33). The experiments are also labor-intensive, time-consuming and tedious. In addition, the number of possibly interacting protein pairs within one cell will be enormous, which makes complete experimental verification impractical. Therefore, computational methods are constantly needed to complement existing experimental approaches. Several prediction studies have been carried out by deriving information from the vast amount of biological data contained in the genomic datasets, such as gene neighborhood (34–36), gene fusion events (37,38), gene co-occurrences or phylogenetic profiles (39–41) and correlated mRNA expression patterns (42,43). In addition, protein interactions can also be extracted from the literature (44–46). A comprehensive overview of these methods can be found elsewhere (47,48). Recently, in order to gain a more comprehensive understanding of the interactome, based on a single probabilistic framework, different genomic features were integrated to make large-scale predictions of protein–protein interactions in yeast (13,49) and human (14). As expected, prediction should be improved with the integration of more independent genomic features even if each one is a weak predictor of protein interactions (50). However, it is known that any delicate dependencies between features can confound the strength of the prediction in these integrated frameworks, although there may be no appreciable statistical dependence between the many possible pairs of these features (51).

Once protein–protein interaction networks have been reconstructed, either experimentally or computationally, they are usually analyzed to relate structural properties of networks with protein properties on a global or local topology view (48,52). In this study, we focused mainly on identifying the structural relationships among members of protein complexes. It is realized that no protein is an isolated island, but instead most seem to function by binding together in complexes (53); many important cellular functions are actually carried out by protein complexes that act as molecular 'machines' (54). Moreover, there may exist a higher-order organization of interacting complexes for the coordination of cellular functions (30). Meanwhile, it has been shown that many complexes in yeast and humans are nearly identical, which provides an understanding that, rather than at the protein level, they are conserved at the machine level during the course of evolution (55).

In this paper, we define a new metric for semantic similarity to score the degree of the functional association between two different proteins by comparing the relative specificity of pairs of GO terms assigned to them in similarity within a GO DAG. As mentioned above, both the CC and BP ontologies and their respective annotations were used in this study. To evaluate the method, an integrated high-quality interaction dataset was applied. Based on the evaluation, a positive and a negative dataset were selected, and then used to assess the four large-scale experiments mentioned above (28–31). The result of the assessment is in agreement with that of the previous studies (32,33). In addition, we used the map reconstructed from the GSPs, which is with highest confidences in positives, to analyze the internal possible interacting relationships among the partners in the MIPS complexes. This reveals that our method may be a little biased and that the predicted map seems to be more comprehensive than those obtained from the other approaches mentioned above. Accordingly, our method may also be applied to the other completely sequenced genomes that are well annotated with the GO schemes, such as the human genome.

MATERIALS AND METHODS

The GO and yeast annotations

Yeast protein annotations were downloaded from the Organelle DB (56) and for the compatibility of computation, the September 2004 release of the GO was used. Organelle DB is the first on-line resource devoted to the identification and presentation of eukaryotic proteins localized to organelles and subcellular structures. In the simple eukaryote *S.cerevisiae* (yeast), Organelle DB collects and presents several large-scale protein localization projects (21,57,58) and the localization data that has been generated piecemeal from independent small-scale studies. Furthermore, to facilitate data interoperability, proteins in Organelle DB have been annotated using the three controlled vocabularies (BP, CC and MF) from the GO consortium.

Data filtering criteria

In order for computational effectiveness and clarity, we excluded the following GO terms from the analysis:

(i) GO terms that are defined as 'biological_process unknown' (GO: 0000004) (including 641 annotations) in BP ontology;
(ii) For the CC ontology, there are six terms descending directly from its root (GO: 0005575), namely, 'cellular_component unknown' (GO: 0008372), 'unlocalized' (GO: 0005941), 'virion' (GO: 0019012), 'immunoglobulin complex' (GO: 0019814), 'extracellular' (GO: 0005576) and 'cell' (GO: 0005623). Only the term 'cell' was used in the analysis because we focused on the proximity of co-localized yeast protein pairs. In addition, we found that there were only eight proteins annotated with the term 'extracellular' and no proteins were assigned to its descendant terms. These eight proteins were also annotated with other terms descending from the term 'cell' (Supplementary Table S1). Therefore, the term 'cell' was then set to be the root of the GO cellular component in this study.

The distributions of GO terms and the respective yeast protein annotations before and after applying the procedure of filtering are listed in Supplementary Table S2. Our analysis was thereafter based on the filtered datasets.

Seven known protein–protein interaction datasets

Seven existing protein–protein interaction datasets were used for validation of our method (D5–7) and for assessment of their accuracy by our predicted interaction dataset (D1–4). They are:

D1–2: datasets 'Gavin' and 'Ho'; both of them consist of binary interactions converted from the data inferred from mass spectrometry of coimmunoprecipitated complexes (30,31) using the spoke model, which has been shown to be more reliable than the matrix model in this case (59,60).

D3–4: datasets 'Ito' and 'Uetz', each from a different independent genome-scale yeast two-hybrid experiment (28,29).

D5: 'MIPS complexes' dataset, which comprises binary interactions converted from MIPS complexes (61) without topological information using a matrix model.

D6: 'MIPS interactions' dataset which is composed of those MIPS physical interactions (61) that have been inferred from small-scale experiments.

D7: 'de Lichtenberg' dataset, which refers to those integrated interactions involved in the processes during the yeast cell cycle (62).

The numbers of proteins and interactions of the seven known protein–protein interaction datasets are listed in Supplementary Table S3.

Relative Specificity Similarity (RSS) of two proteins annotated in a GO

Since each GO is structured as a DAG, wherein one term is a child of one or multiple parents, and child terms are instances (is-a relationship) or components (part-of relationship) of parent terms, there is often more than one path from a GO term up to the topmost level of the GO, namely, the root term of the GO. In this paper, the topmost level of a GO indicates the root term 'biological_process' (GO: 0008150) of the BP ontology or the term 'cell' (GO: 0005623) of the CC ontology. As stated in Wu's definition, the collection of paths with each one corresponding to a complete trace from the concerned GO term to the root term of the GO can be represented as a graph induced from the concerned term (17). For a given GO, let term$_i$ and term$_j$ be two terms, Paths(term$_i$) and Paths(term$_j$) be the paths in the graphs induced from term$_i$ and term$_j$ respectively, and dist(u, v) be the number of edges along the shortest path between term u and term v, so that its value equals zero if u and v are the same term. Three different configurations may exist for two terms, term$_i$ and term$_j$ from a given GO (Figure 1). In each configuration, the RSS of two GO terms consists of three different components. They are denoted α, β and γ, respectively. Component α is defined in formula 1; it measures how specific the MRCA of the two terms is according to the structure of the GO and is equivalent to the definition of S in Wu's work (17),

$$\alpha = \max_{\substack{\text{path}_m \in \text{Paths}(\text{term}_i),\\ \text{path}_n \in \text{Paths}(\text{term}_j)}} \left\{ \begin{matrix} \text{the number of common terms} \\ \text{between path}_m \text{ and path}_n \end{matrix} \right\} - 1 \qquad \mathbf{1}$$

Component β measures how relatively general term$_i$ and term$_j$ are in the GO and is defined in formula 2. The generality of a term is defined as the minimum distance between the term and all of the leaf terms descending from it. Leaf terms in a GO are those terms without any descendant. Obviously, the larger the distance, the more general is a term.

$$\beta = \max\{\min_{u \in U}\{\text{dist}(\text{term}_i, u)\}, \min_{v \in V}\{\text{dist}(\text{term}_j, v)\}\} \qquad \mathbf{2}$$

where U = {all leaf nodes descending from term$_i$} and V = {all leaf nodes descending from term$_j$}.

Component γ measures the local distance between two terms relative to the MRCA and is defined as follows:

$$\gamma = \text{dist}(\text{MRCA}, \text{term}_i) + \text{dist}(\text{MRCA}, \text{term}_j) \qquad \mathbf{3}$$

If γ is small, it implies term$_i$ and term$_j$ share much similarity locally relative to the MRCA.

Then, the RSS between two terms of a given GO, term$_i$ and term$_j$ can be quantified by combining α, β and γ together in formula 4,

$$\text{RSS}(\text{term}_i, \text{term}_j) = \frac{\text{maxDepth}^{\text{GO}}}{\text{maxDepth}^{\text{GO}} + \gamma} \cdot \frac{\alpha}{\alpha + \beta} \qquad \mathbf{4}$$

where maxDepth$^{\text{GO}}$ is the maximum distance from the root term of the GO to the leaf terms (i.e. the number of edges along the longest path in the GO). From the definition, the values of RSS are between 0 and 1. Clearly, RSS = 0 (α = 0) indicates that the MRCA of term$_i$ and term$_j$ is the root of the GO, which means that the two terms share no commonality in describing protein properties; on the other hand, RSS = 1 (γ = 0 and β = 0) indicates that term$_i$ and term$_j$ are the

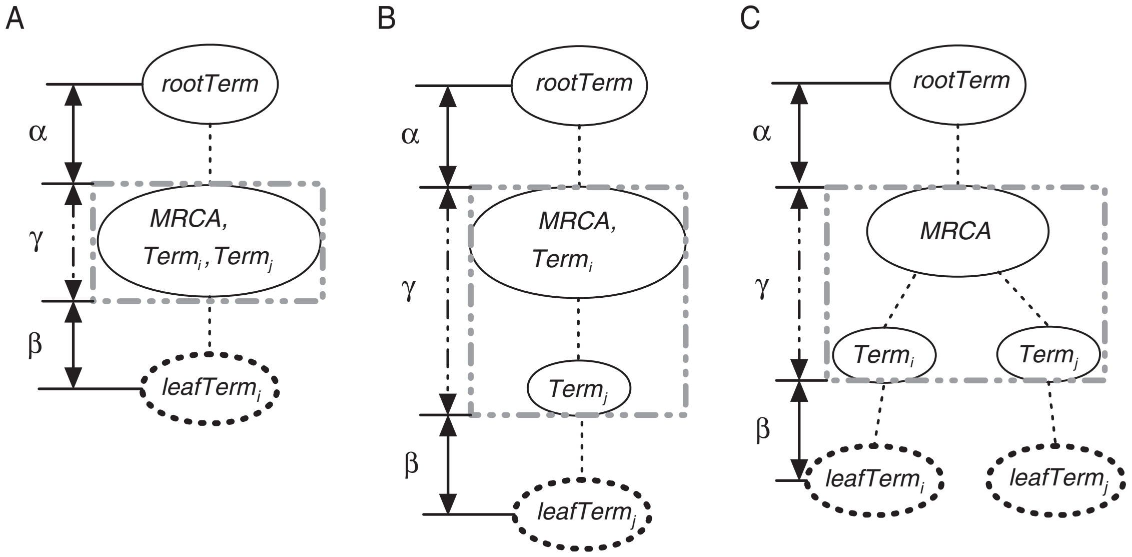


Figure 1. Three different configurations describing two terms in a given DAG. MRCA which is called the most recent common ancestor of term$_i$ and term$_j$, represents the most specific of all common ancestors of the term pair. (**A**) Two terms overlap; (**B**) term$_j$ is a descendant of term$_i$. So term$_i$ is also their MRCA; (**C**) No path exists between the two terms. Dashed lines indicate that there may be more than one path between two terms. Similarly, dashed nodes represent one or multiple leaf terms descending from a 'parent' term.

same leaf term, which means that the two terms are most specific in describing protein attributes.

Based on the definition of RSS between the two terms, we can formalize a metric for measuring the relationship strength, including the functional association or the location proximity, between two different proteins annotated in the GO. Let P and Q be two proteins of interest, and terms(P) and terms(Q) the sets of all the GO terms assigned to protein P and Q, respectively. We define the relationship strength between P and Q, $\mathrm{RSS}^{\mathrm{GO}}(P, Q)$, as the maximum RSS of all possible term pairs from terms(P) and terms(Q), respectively, namely,

$$\mathrm{RSS}^{\mathrm{GO}}(P,Q) = \max_{\substack{u \in \mathrm{terms}(P) \\ v \in \mathrm{terms}(Q)}} \{\mathrm{RSS}(u,v)\} \qquad \mathbf{5}$$

Statistical significance of protein pairs falling in various levels of RSS values

In order to determine whether the assignment of protein pairs into categories with different RSS values is statistically significant and to minimize the systematic biases intrinsically raised from the structure of a GO, a Z-score analysis was applied. Firstly, the [0,1] interval was equally divided into 10 categories, namely, $\{(0.1 \times i,\ 0.1 \times (i + 1)],\ i = 0,1,\ldots,9\}$, plus the other one with RSS equal to 0, which indicates that the MRCA of all term pairs respectively assigned to two proteins is the root of the GO. Then, Z-score values, defined as $(\#pairs_{\mathrm{annotated}} - \#pairs_{\mathrm{random}})/SD_{\mathrm{random}}$, were calculated from the number of protein pairs from a given GO annotation (BP or CC) and random annotation based on the GO DAG. For computational simplicity and without loss of generality, we randomly assigned each of all distinct proteins to one term of the GO and then calculated the number of protein pairs falling in an RSS category. This process was repeated 1000 times (we have also tested with 5000 times and a similar result was found) and afterwards the corresponding average numbers ($\#pairs_{\mathrm{random}}$) and their standard deviations (SD_{random}) were calculated. As we know, the larger the Z-score value, the less probable it is that the relationship strength of a pair of proteins is due to chance from the structure of ontology. Therefore, the Z-score value for each RSS category indicates the confidence of the relationship strength of protein pairs measured by our method.

RESULTS

Distributions of all pairs of annotated proteins according to their RSS values

In all, 3832 proteins are assigned to one or more of 265 CC terms in Organelle DB, which produces 5722 annotations and 7 340 196 (3832×3831/2) pairs of different proteins in the CC ontology. Similarly, there are 5574 annotations and 5 092 836 protein pairs in the BP ontology (Supplementary Table S2). For a given GO DAG, an RSS is assigned to each pair of different proteins according to formula 5. The distributions of all annotated protein pairs in various RSS categories for the two GOs are shown in Figure 2, where the [0,1] interval of RSS values is split into 11 categories (see Materials and Methods). Each blue line represents the cumulative rate of protein pairs along with the 11 categories. In order to draw the statistical significance of protein pairs falling in various levels of RSS values for the GO, Z-score values for each of the 11 categories were calculated (see Materials and Methods for details). As shown in Figure 3, in each of the two RSS categories, (0.9, 1] and (0.9, 0.8], the number of pairs of proteins annotated on either the CC or BP ontology is more than 142 standard deviations greater than the mean number with randomized annotations. When the RSS values are equal to or less than 0.3 for the CC ontology and are equal to or less than 0.4 for the BP

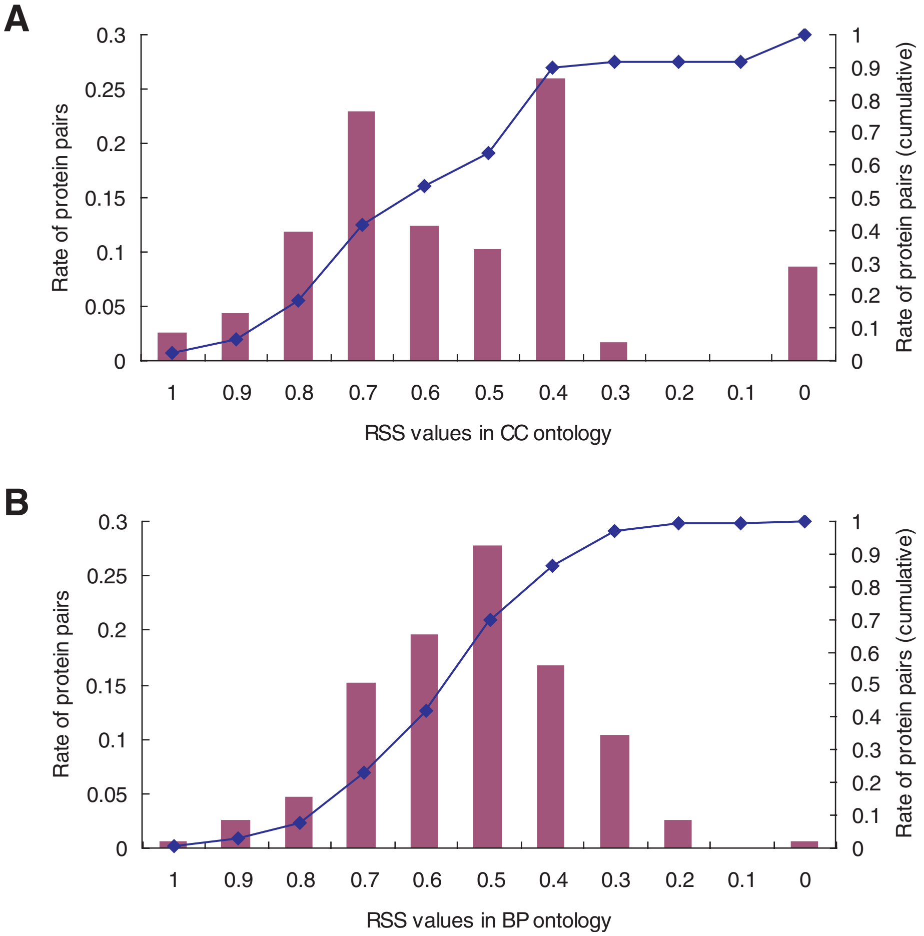


Figure 2. Distributions of the annotated protein pairs with various RSS values in the CC (**A**) and BP (**B**) ontologies. The [0,1] interval of the *x*-axis is equally divided into 11 categories, as defined in Materials and Methods. The histogram (relating to the left *y*-axis) indicates the rate of protein pairs falling into the specified ranges of RSS values, and the blue line (relating to the right *y*-axis) shows the cumulative rate of protein pairs.

ontology, no statistical significance could be detected. The *Z*-score values are less than 1.5, indicating that most, if not all, of these pairs may not be functionally associated.

Based on the distribution analyses of *Z*-scores for CC and BP ontologies, the 11 categories of RSS^{CC} could roughly be divided into three groups with high confidence (H, $0.8 < RSS^{CC} \leq 1$), medium confidence (M, $0.3 < RSS^{CC} \leq 0.8$) and low confidence (L, $0 \leq RSS^{CC} \leq 0.3$). Like RSS^{CC}, RSS^{BP} can also be split into three groups, with high confidence (H, $0.8 < RSS^{BP} \leq 1$), medium confidence (M, $0.4 < RSS^{BP} \leq 0.8$) and low confidence (L, $0 \leq RSS^{BP} \leq 0.4$). Therefore, there are nine (3×3) data segments (DSs) (a total of 5 010 195 protein pairs encompassing 3166 proteins) with different combinations of confidences according to the subdivisions of both RSS values, for example, a DS which consists of protein pairs with high-confidence RSS^{CC} (H) and medium-confidence RSS^{BP} (M) is called HM DS (Figure 4A).

Gold standard positive and negative protein interaction datasets

How likely is one pair of proteins from each of these nine DSs to interact with each other physically? In order to address this issue, the union of three existing protein–protein interaction datasets—binary interactions in a matrix model from the MIPS complexes, the MIPS small-scale physical interactions, and the integrated interactions by de Lichtenberg *et al.* (see Materials and Methods)—was chosen as trusted, and thus, is called 'valid experimental interactions (VEIs)'. The MIPS complexes and the MIPS physical interactions are often used as or as part of 'gold standard positives' to validate various prediction methods (13,51,63) and are also used to assess high-throughput interaction datasets (32,33). There are 11 041 unique binary interactions among 1472 proteins in our VEIs, and Supplementary Figure S1 shows the distribution of interactions among the three datasets. The distribution of VEIs among these nine DSs is shown in Figure 5A. Interestingly, 78% (8620 out of 11 041) of interactions fall into the HH DS, suggesting that the HH DS may contain most, if not all, of yeast protein–protein interactions; whereas 0.06% (7 out of 11 041) of interactions are in the LL DS, suggesting that protein pairs in LL DS seem to be much less likely to interact physically. Similar to the process mentioned above, we applied the analysis of statistical significance using *Z*-score values calculated in each DS (Figure 5B). As a result, we could roughly classify the nine DSs into two groups, one called the 'positive dataset' (3101 proteins and 152 944 interactions) including two DSs (HH and MH) with *Z*-score values larger

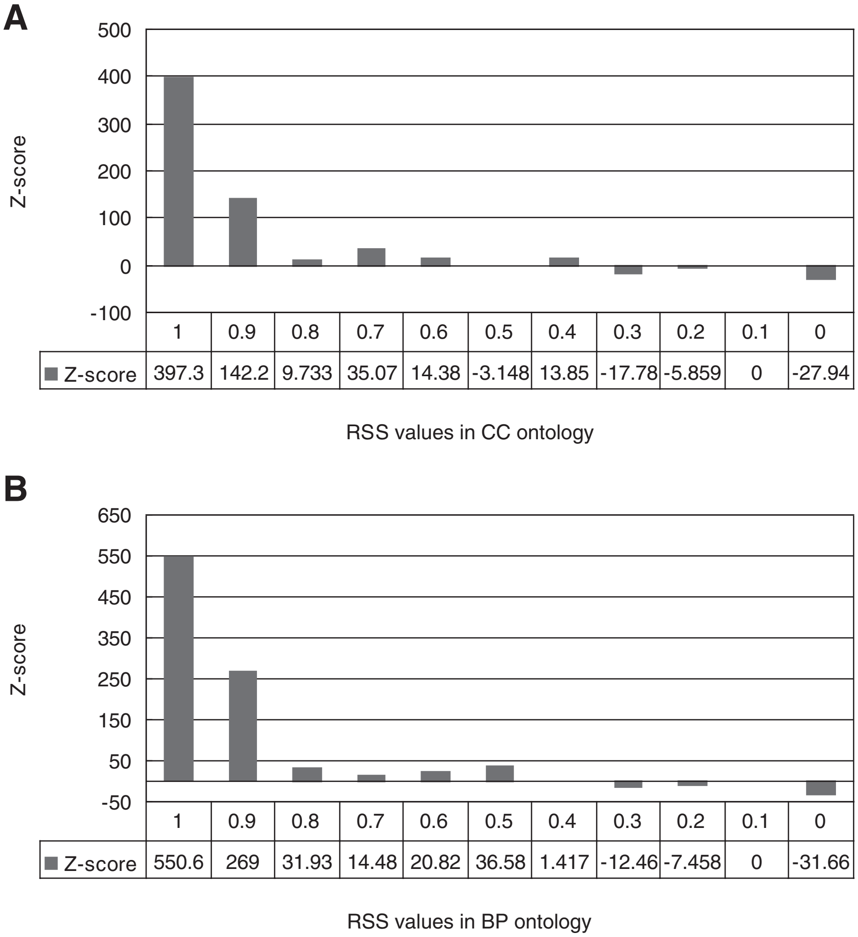


Figure 3. Statistical significance of the quality scoring system. For a given GO, which is either CC (**A**) or BP (**B**), *Z*-score values, defined as $\#pairs_{\mathrm{annotated}}-\#pairs_{\mathrm{random}})/SD_{\mathrm{random}}$, were calculated from the number of protein pairs from the given GO annotation and random annotation based on the GO DAG. The [0,1] interval of the *x*-axis is divided equally into 11 RSS categories, as defined in Materials and Methods.

than 320, and the other called the ‘negative dataset’ (3166 proteins and 4 857 251 interactions) including the remainder of the DSs (HL, LL, LM, MM, HL, LH and HM) with *Z*-score values ranging from −30 to 42. In particular, the HH DS, whose *Z*-score values reaches 4019, was selected as a gold standard positive dataset (GSPs; 2259 proteins and 40 753 interactions), whereas two DSs (ML and LL), whose *Z*-score values are lower than −23, were combined as a gold standard negative dataset (GSNs; 3165 proteins and 1 460 378 interactions) in this study (Figure 4).

Assessment of four known genome-scale experimental datasets

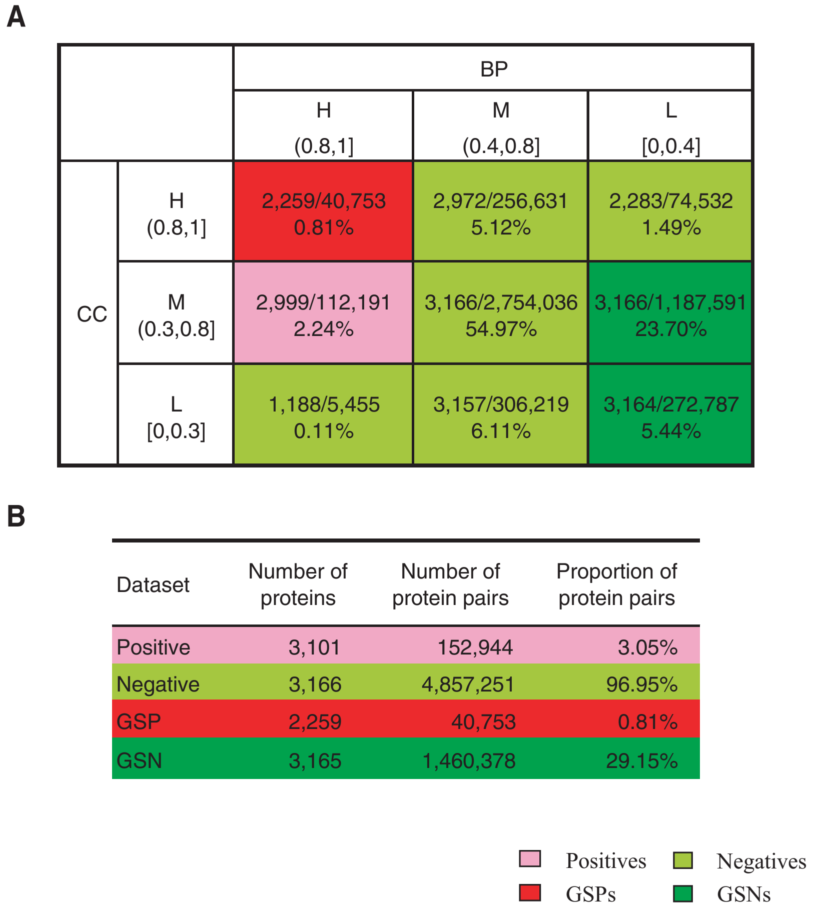
Four known genome-scale protein interaction datasets (D1–4) were assessed using our positives and negatives (Supplementary Figure S2A). The coverage of two ‘pull-down’ interaction datasets (30,31) in positives is 60% (1404/2324 for the Gavin dataset) and 23% (402/1766 for the Ho dataset), respectively, while the coverage of two genome-scale Y2H interaction datasets (28,29) in positives reaches 35% (126/356 for the Uetz dataset) and 14% (190/1392 for the Ito dataset), respectively. It is worth noting that, assessed by our negative dataset, Gavin, Uetz, Ho and Ito datasets contain very high proportions of false-positive interactions, of 40% (920/2324), 65% (230/356), 77% (1364/1766) and 86% (1202/1392), respectively.

The four genome-scale interaction datasets were also assessed using GSPs and GSNs (Supplementary Figure S2B) and a similar result was obtained. For the Gavin dataset, 42% (984/2324) are covered by GSPs, while only 6% (141/2324) are present as false positives. For Uetz, Ho and Ito datasets, 14% (51/356), 11% (188/1766) and 5% (72/1392) are respectively covered by GSPs, while 14% (49/356), 15% (271/1766) and 27% (371/1392) are found to be false positives, respectively.

Consequently, it appears that the Gavin dataset discovers true interactions at a larger coverage and contains a lower proportion of false-positive interactions, whereas Uetz, Ho and Ito datasets have smaller coverage of true interactions and probably populated by more false positives. These observations are in agreement with the assessment result reported in the previous studies (32,33).

Initial analysis of the topologies of the MIPS complexes with the predicted network

In order to minimize the error rate of the predicted interactions, only the highest-confidence interaction dataset called



A

		BP H (0.8,1]	BP M (0.4,0.8]	BP L [0,0.4]
CC	H (0.8,1]	🟥 2,259/40,753 0.81%	🟩 2,972/256,631 5.12%	🟩 2,283/74,532 1.49%
	M (0.3,0.8]	🟪 2,999/112,191 2.24%	🟩 3,166/2,754,036 54.97%	🟩 3,166/1,187,591 23.70%
	L [0,0.3]	🟩 1,188/5,455 0.11%	🟩 3,157/306,219 6.11%	🟩 3,164/272,787 5.44%

B

Dataset	Number of proteins	Number of protein pairs	Proportion of protein pairs
Positive	3,101	152,944	3.05%
Negative	3,166	4,857,251	96.95%
GSP	2,259	40,753	0.81%
GSN	3,165	1,460,378	29.15%

□ Positives
□ Negatives
■ GSPs
■ GSNs

Figure 4. Nine data segments (DSs) with different confidences related to CC and BP ontologies (**A**) and the selection of positives and negatives, as well as GSPs and GSNs (**B**). Nine DSs contain 5 010 195 protein pairs encompassing 3166 proteins in total. Each DS is labeled as number of proteins/number of protein pairs and the proportion of protein pairs covered by the DS (A). Similar labels are also shown for positives and negatives, as well as GSPs and GSNs (B). Three confidence levels of the protein pairs annotated in the CC ontology are high (H; RSS^{CC} in (0.8, 1]), medium (M; RSS^{CC} in (0.3, 0.8]) and low (L; RSS^{CC} in [0, 0.3]); while three levels of protein pairs annotated in BP ontology are high (H; RSS^{BP} in (0.8, 1]), medium (M; RSS^{BP} in (0.4, 0.8]) and low (L; RSS^{BP} in [0, 0.4]). The nine DSs are divided into two parts, which are positives (HH and MH in rose color) and negatives (the remaining seven DSs in lime). The gold standard positive dataset (GSPs; HH in red) is that part of the positives with the highest confidence while the gold standard negative dataset (GSNs; ML+LL in green) is that part of the negatives with the lowest confidence.

GSPs was used here in the network analysis. Based on the network reconstructed from GSPs, we could proceed to analyze it using various approaches and algorithms of graph theory to relate its structural properties to protein functions. A good overview of these analyses can be found in Xia *et al.* (48). Here, we focus on the identification of the topology of each of the MIPS complexes using our predicted network. There are 40 753 interactions encompassing 2259 proteins in GSPs (Figure 4), and the whole interaction network derived from the dataset consists of 16 connected components (Table 1, see also Supplementary Figure S3). Among these 16 components, the largest one (connected component ID: 1 as listed in Table 1) contains 30 899 interactions among 2093 proteins, and seven (connected component ID: 10–16 as listed in Table 1) are each composed of only one interaction between two proteins. In addition, out of 600 periodically expressed ('dynamic') proteins from de Lichtenberg *et al.* (62), we found 319 ones in our nine DSs and 228 in GSPs (Table 1, see also Supplementary Figure S3 where dynamic proteins are in red). Within interaction networks, a protein complex is ideally identified as a 'complete subgraph' where every pair of a complex's members tends to interact with each other (13). Therefore, various clustering techniques are suggested to detect protein complexes (64,65). However, this rarely happens in reality, such as in the Arp2/3 complex in yeast (66). Here, we are interested in the analysis of the structures of the MIPS complexes based on the predicted network. After excluding 50 homodimer complexes, 214 MIPS complexes with at least two distinct members were analyzed. There are 120 complexes, each with all members found in the predicted interaction network. Consequently, 76 out of the 120 complexes are each interconnected in the network, including 71 within the largest connected component (connected component ID: 1) and the remaining five in the other four small connected components (connected component ID: 3, 5, 9 and 11) (Supplementary Table S4). Such a topology of a complex from the 76 ones is called a connected subgraph. Very interestingly, out of the 120 complexes, we found 27 ones that each splits into two connected subgraphs in the largest connected component. Seven of them can be linked by at least one path of

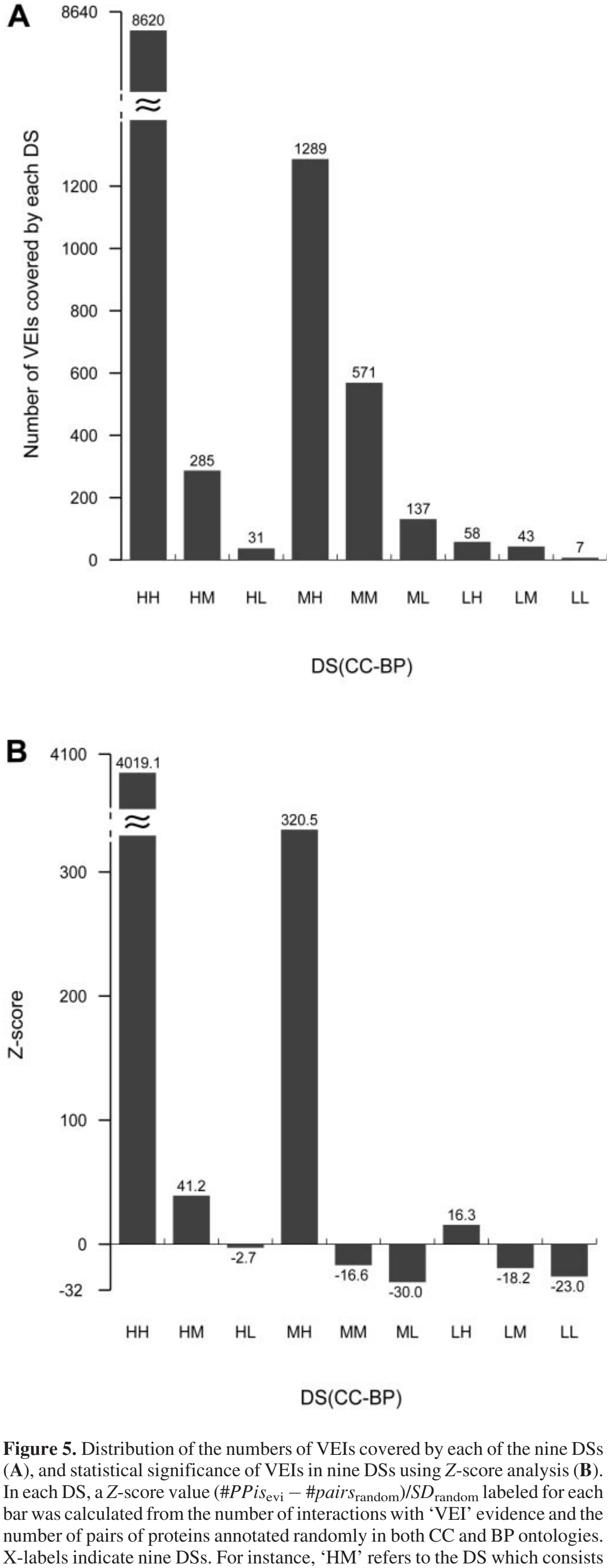


Figure 5. Distribution of the numbers of VEIs covered by each of the nine DSs (**A**), and statistical significance of VEIs in nine DSs using *Z*-score analysis (**B**). In each DS, a *Z*-score value $(\#PPis_{\text{evi}} - \#pairs_{\text{random}})/SD_{\text{random}}$ labeled for each bar was calculated from the number of interactions with 'VEI' evidence and the number of pairs of proteins annotated randomly in both CC and BP ontologies. X-labels indicate nine DSs. For instance, 'HM' refers to the DS which consists of protein pairs with high-confidence RSS^{CC} (H) and medium-confidence RSS^{BP} (M).

Table 1. Numbers of proteins and interactions of 16 connected graphs found in the predicted interaction network constructed from the GSP dataset

Connected component ID	Number of proteins (dynamic)[a]	Number of interactions
1	2093 (228)	36 899
2	85	3570
3	20	100
4	15	95
5	15	59
6	6	11
7	5	6
8	3	3
9	3	3
10	2	1
11	2	1
12	2	1
13	2	1
14	2	1
15	2	1
16	2	1
Total	2259 (228)	40 753

[a] The figure in brackets indicates the number of dynamic interacting proteins identified in the connected component.

two interactions, with one carrying a highest RSS value (=1, either for the CC or for the BP) (Supplementary Table S5), indicating that the proteins along the paths, which are not members of the seven complexes might be biologically related to the function of the complexes.

Analysis of the five complexes in the four small connected components. Figure 6 shows five complexes each with all members forming a connected subgraph in the four small connected components. They are Aut2p/Aut7p complex (the MIPS identifier: 260.100), dynactin complex (140.30.30.30), F0/F1 ATP synthase (complex V) (420.50), succinate dehydrogenase complex (complex II) (420.20) and cytochrome bc1 complex (420.30).

Interestingly, in the mapped F0/F1 ATP synthase complex (Figure 6), we found two nonmember proteins, namely, STF1 and STF2. Both are ATPase stabilizing factors. In particular, STF1 stabilizes and facilitates the formation of the complex between mitochondrial ATP synthase and its intrinsic inhibitor protein (67), while STF2 binds to F0-ATPase and facilitates binding of inhibitor and a 9 kDa protein to F1-ATPase. In the mapped succinate dehydrogenase complex, nonmember NDI1 is an NADH2 dehydrogenase (ubiquinone) (68), while nonmember CYT, which is a member of the cytochrome bc1 complex (Figure 6), functions as an electron transporter and transfers electrons within CoQH2-cytochrome *c* reductase complex activity (69). Therefore, both of them are functionally related to the succinate dehydrogenase complex in terms of cellular processes and biological functions. These findings have two implications. If not new members of a complex, these identified nonmember proteins might interact with the F0/F1 ATP synthase (such as STF1 and STF2) and the succinate dehydrogenase complex (such as NDI1 and CYT). On the other hand, two complexes might interact with each other somewhere, such as the succinate dehydrogenase complex and the cytochrome bc1 complex via the mediator CYT (Figure 6).

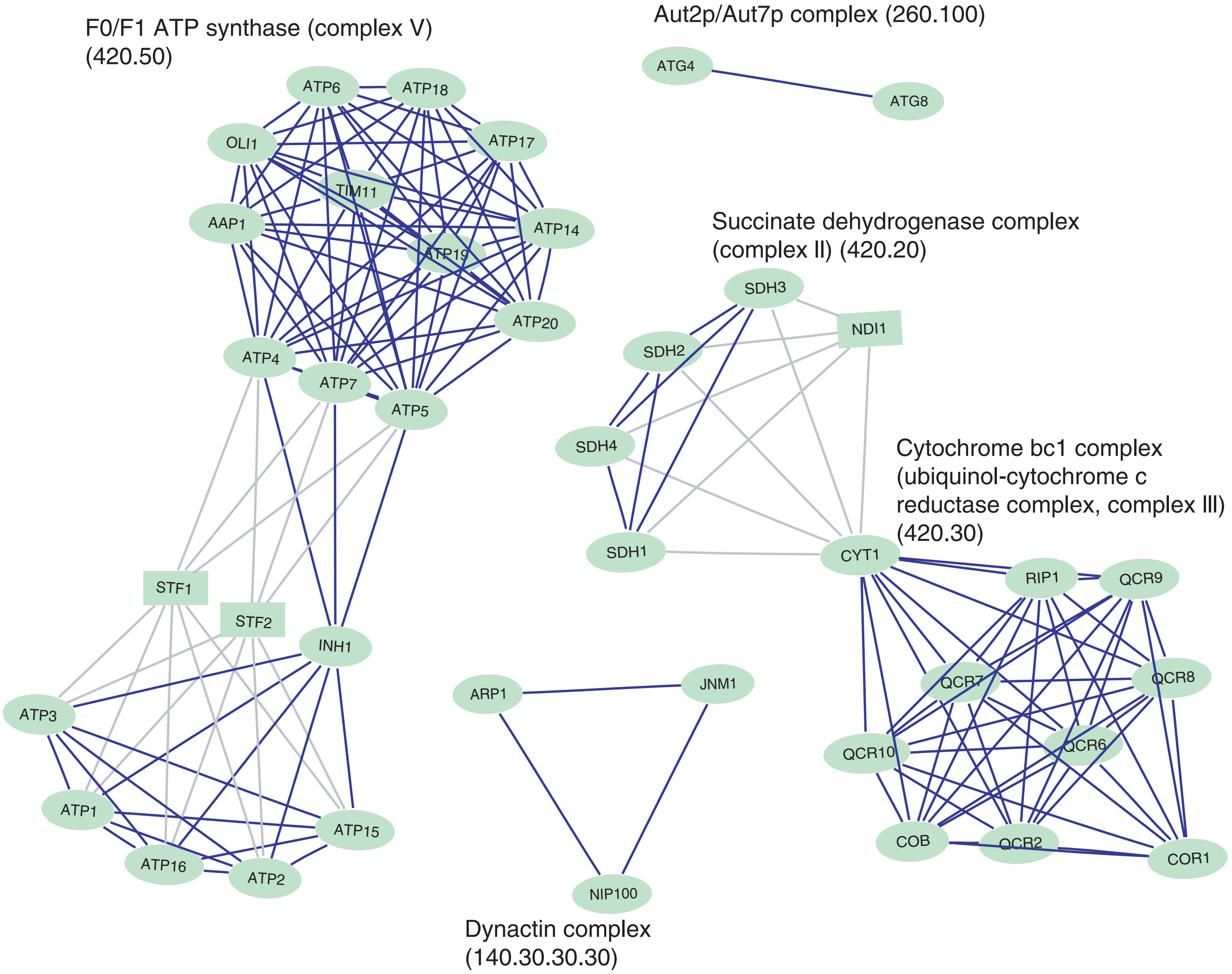


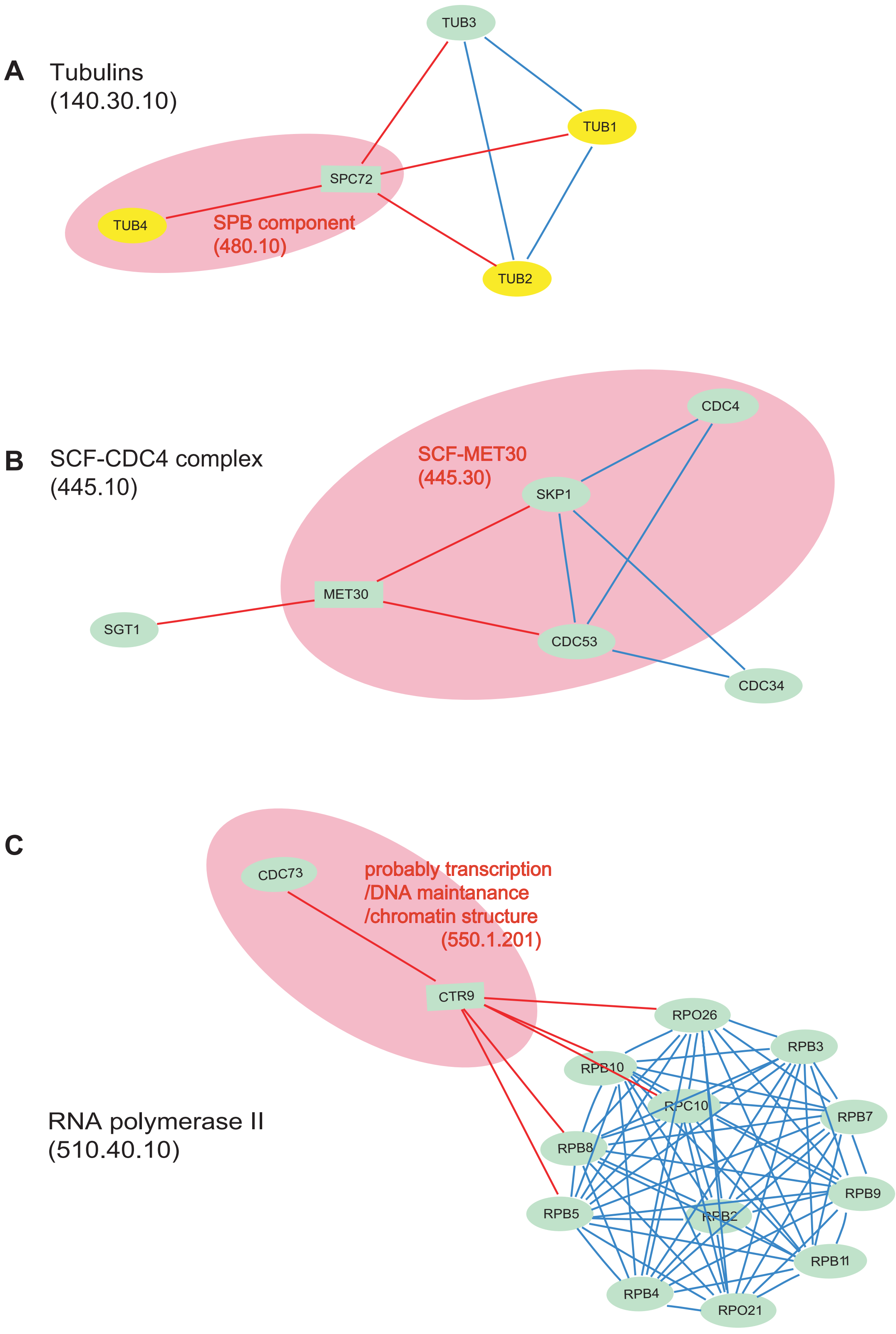
Figure 6. Five complexes each with all members forming a connected subgraph. They are in four small connected components (connected component ID: 3, 5, 9 and 11) of the interaction network constructed from GSPs. Proteins as members, and the nonmember proteins of the five given complexes in the connected components are shown in ellipse and box nodes, respectively. Nodes indicated in dark sea-green represent the static proteins. Interactions within a complex and across two complexes are shown in blue and in gray, respectively. The description and identifier of each MIPS complex are labeled beside the complex.

Analysis of the split complexes. For the mapped complexes split into two connected subgraphs, we searched those paths of two interactions where at least one carries a highest RSS value (=1, either for the CC or for the BP) linking the two connected subgraphs. As a result, we found seven split complexes could be connected as a whole by adding one or multiple such paths (Supplementary Table S5) and five of them are illustrated in Figure 7. Interestingly, as shown in Figure 7A–C, the nonmember proteins in the paths in the three complexes are also members of other complexes. In Figure 7A, SPC72 is found in the SPB components complex (480.10) and its N-terminal domain interacts with the Tub4p complex on the cytoplasmic side of the SPB (70). In Figure 7B, CDC34 and the F-box protein MET30 are required for degradation of the Cdk-inhibitory kinase Swe1, and both of them as well as CDC53 and SKP1 have been found in the SCF-MET30 complex (445.30) (71). In Figure 7C, CTR9 is required for G1 cyclin expression (72) and is also found in another complex that includes CDC73 (30).

Figure 7D shows the cytochrome *c* oxidase (complex IV) (420.40) whose two connected subgraphs are linked by 20 paths in our predicted network. Ten of these 20 paths are all through the protein PET100, while the other 10 paths are all through another protein COX20. Interestingly, both PET100 and COX20 are found to be essential for the assembly of this complex. PET100 is required for the assembly of yeast cytochrome *c* oxidase (420.40) into an active holoenzyme (73–75), whereas COX20 acts as a membrane-bound chaperone necessary for the cleavage of pCox2p (the subunit 2 precursor) and for interaction of the mature protein with other subunits of cytochrome *c* oxidase in a later step of the assembly process (76).

As shown in Figure 7E, LRP1 is implicated in both nonhomologous DNA end joining and homologous recombination (77), and therefore the complex DNA ligase IV (510.180.30.20) might have one more partner (LRP1) than identified currently. This is expected to provide a clue for further experimental validation.

In summary, 35% (76/214) of the MIPS complexes are completely identified from our interaction network. Furthermore, 13% (27/214) of complexes, all of whose partners are in the largest connected component of the network, are split into two connected subgraphs each. Five such split complexes are linked by at least one reliable path of two interactions, and the nonmember proteins in the paths actually physically interact with the complexes concerned. We also predict that the



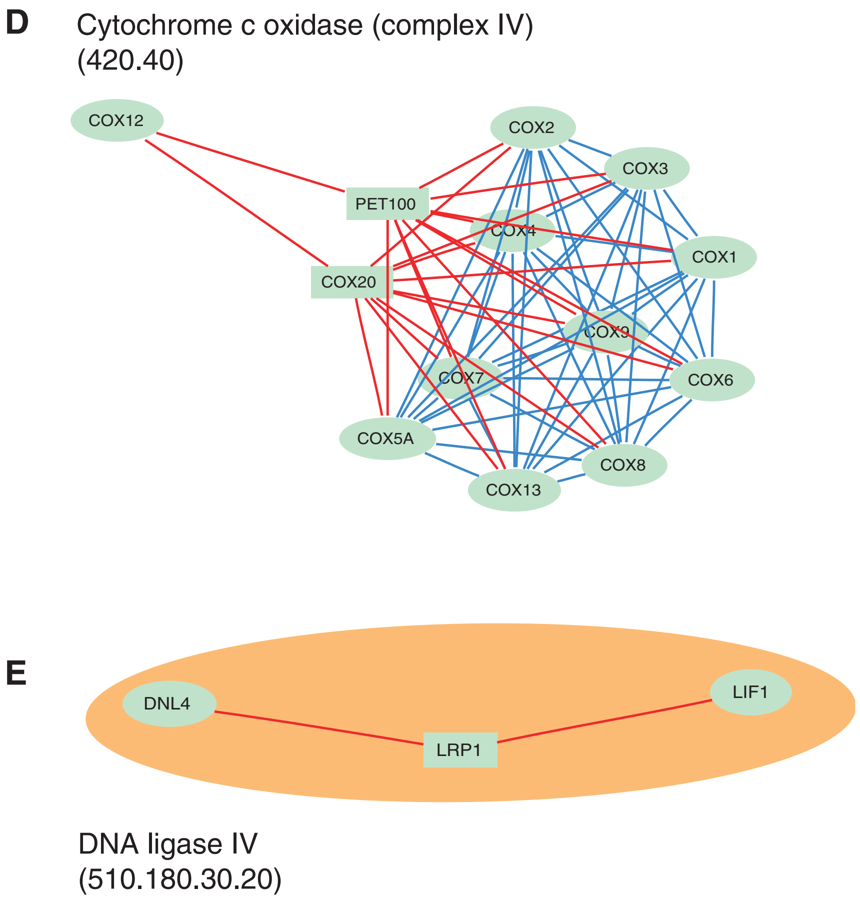


Figure 7. Five complexes whose members are in one connected component of the network and each of which splits into two connected subgraphs that can be linked by at least one path of two interactions with one carrying a highest RSS value (=1, either for the CC or for the BP). Nonmember proteins in the paths are considered to be functionally associated with the concerned complexes. As shown in A, B and C, the proteins in one pink region are also members of another complex (other than the given five) whose name and MIPS identifier are labeled in red. As shown in D, two nonmember proteins are essential for the compiling of the complex. Furthermore, as shown in E, the two-member complex DNA ligase IV might have one more partner (LRP1) than identified currently and therefore the three proteins are grouped in an orange region. The descriptions and MIPS identifiers of the five complexes and others with nonmember proteins are indicated in black and red, respectively. Proteins as members and nonmembers of the five complexes are respectively shown in ellipse and box nodes. All proteins whose standard names are labeled in the nodes are either dynamic (in yellow) or static (in dark sea-green). Interactions in the complexes and across complexes are shown as blue and red edges, respectively.

complex DNA ligase IV might contain one more partner, which deserves experimental confirmation (Figure 7E).

DISCUSSION

We present here a new method, which is based on semantic similarity measures, to reconstruct the map of yeast protein–protein interactions by mining the knowledge of functional associations from the GO-based annotations. As a result, a positive and a negative datasets containing 152 944 and 4 857 251 interactions, respectively were derived. Moreover, we compiled GSPs with high confidence and GSNs with low confidence of yeast protein–protein interactions, containing 40 753 and 1 460 378 interactions, respectively. It is estimated that the number of actual interactions in yeast ranges from ~30 000 to 100 000 (32,53,59). Interestingly, the number of protein–protein interactions in GSPs falls into this range. In particular, our GSPs contain 78% of those interactions in VEIs, which indicates that GSPs may be less biased than other published datasets. Furthermore, the number of protein pairs in GSNs is 35.8 times larger than the number in GSPs, which is in accordance with the expectation that the estimated number of non-interacting protein pairs is several orders of magnitude higher than the number of positives (78).

As regards the application of GSPs and GSNs, it has been found that, in addition to the method of choice of highly reliable interactions (positives), how unbiased negative examples are chosen also has a strong effect on the performance of any of the supervised machine learning methods for prediction of protein–protein interactions. Until now there have been several strategies to select negative datasets for detecting protein–protein interactions, such as choosing random pairs of interacting proteins (79) and selecting the pairs of proteins that are known to be localized in different cellular components (13,63). Since it is possible that two proteins localized distinctly (e.g. in the nucleus and cytoplasm) may sometimes physically interact (19), in this study, protein pairs both involved in weakly-related or unrelated biological processes and localized in different cellular components are considered for and compiled into our GSNs. Thus, it seems that the resulting GSN dataset is less biased compared with those

Table 2A. Seven VEIs in the LL DS

Standard name	Systematic name	Standard name	Systematic name	Interaction dataset[a]	Evidence (complex or PMID)[a]
CAF4	YKR036C	MOB1	YIL106W	MIPS complexes	CCR4 complex (510.190.110)
CAF16	YFL028C	MOB1	YIL106W	MIPS complexes	CCR4 complex (510.190.110)
				MIPS interactions	9 528 782
FIR1	YER032W	SPC24	YMR117C	MIPS interactions	9 520 439
FIR1	YER032W	SPO12	YHR152W	MIPS interactions	10 564 265
CBP3	YPL215W	PRP11	YDL043C	MIPS interactions	9 207 794
CBP3	YPL215W	SNP1	YIL061C	MIPS interactions	9 207 794
PRP11	YDL043C	YDR131C	YDR131C	MIPS interactions	9 207 794

Each of them is either from the MIPS physical interaction dataset or from a matrix model for the MIPS curated complexes.

[a] The interaction may be from the MIPS complex or MIPS physical interaction dataset. For the MIPS complex, the complex description as well as its ID in MIPS is represented. For the MIPS physical interaction, the PMID of the literature is represented.

constructed using either location proximity criteria or a randomization strategy.

There are several caveats of our method. First, obviously, the quality of the predicted positives and negatives is constrained by the accuracy of the yeast GO annotations or by our design approach. For instance, we have found that there are seven VEIs in the LL DS (Table 2A) and the protein annotations in the CC and BP ontologies are listed in Table 2B. Six interactions are inferred from small-scale experiments, except the one between CAF4 and MOB1, which is inferred from a matrix model that represents a MIPS complex (the CCR4 complex, 'MIPS identifier: 510.190.110') of unknown topology. It is anticipated that our method will be an avenue for future work following more accurate GO annotations. Second, the elimination of GO terms annotated as 'unknown biological process' (641 annotations) or 'unknown cellular component' may cause some true-positive interactions missing from our predicted network, in the case that the interacting partners are unknown in the biological process or cellular component. This problem can be solved in part as the biological knowledge of these 'missing' proteins accumulates in the future. Third, it has been demonstrated that protein–protein interaction networks in several eukaryotic organisms contain significantly more self-interacting proteins (homodimers) than would be expected if such interactions randomly appeared in the course of evolution (80). However, our method fails in predicting such interactions between the same proteins. Therefore, 50 MIPS homodimer complexes each containing only one protein are excluded from the analysis in this study. Finally, although the strength of relationship between two proteins from our predicted interaction dataset shows high significance, we must note that all predicted interactions should be validated for their functionality by experimental approaches.

As we know, cellular functions are likely to be carried out in a manner of functional modules that often encompass protein complexes (23). As analyzed in this study, out of 120 complexes whose partners are all mapped in our network, 22% (26/120) contain one or multiple dynamic proteins. These dynamic complexes may be just-in-time synthesized, such as nucleosomal protein complex (320) (Supplementary Table S4), or just-in-time assembled, such as replication complex (410.35) (Supplementary Table S4) (62). Functional modules can include transient regulated elements of a relatively distinct process, for example, various transcriptionally regulated cyclins and inhibitors associated with the Cdc28p module at their specific time of synthesis during the yeast mitotic cell cycle (62). Therefore, it is more interesting to do analysis on the temporal properties (such as modules) of the predicted networks rather than on static topological properties, because such studies will provide a basis for further prediction of detailed gene functions and prediction of biological pathways (81).

Table 2B. CC and BP annotations of the proteins for the seven VEIs in the LL DS

Standard name	GO type	GO term	GO ID
CAF16	CC	Cytoplasm	GO: 0005737
	CC	CCR4-NOT complex	GO: 0030014
	BP	Regulation of transcription, DNA-dependent	GO: 0006355
CAF4	CC	CCR4-NOT complex	GO: 0030014
	BP	Regulation of transcription, DNA-dependent	GO: 0006355
CBP3	CC	Mitochondrial membrane	GO: 0005740
	BP	Protein complex assembly	GO: 0006461
FIR1	CC	Bud neck	GO: 0005935
	BP	mRNA polyadenylylation	GO: 0006378
MOB1	CC	Bud neck	GO: 0005935
	BP	Regulation of exit from mitosis	GO: 0007096
	BP	Protein amino acid phosphorylation	GO: 0006468
PRP11	CC	snRNP U2	GO: 0005686
	BP	Spliceosome assembly	GO: 0000245
SNP1	CC	Commitment complex	GO: 0000243
	CC	snRNP U1	GO: 0005685
	BP	Nuclear mRNA splicing, via spliceosome	GO: 0000398
SPC24	CC	Condensed nuclear chromosome kinetochore	GO: 0000778
	CC	Condensed nuclear chromosome, pericentric region	GO: 0000780
	BP	Chromosome segregation	GO: 0007059
	BP	Microtubule nucleation	GO: 0007020
SPO12	CC	Nucleus	GO: 0005634
	CC	Nucleolus	GO: 0005730
	BP	Mitotic cell cycle	GO: 0000278
	BP	Meiosis I	GO: 0007127
	BP	Regulation of exit from mitosis	GO: 0007096
YDR131C	CC	Ubiquitin ligase complex	GO: 0000151
	BP	Ubiquitin-dependent protein catabolism	GO: 0006511

In summary, protein–protein interaction networks in various organisms are increasingly becoming the focus of understanding the functional organization of the proteome. Although only the yeast genome is demonstrated in this study, our method is expected to be applied to other completely sequenced genomes with high-quality annotations based on the GO or other ontologies, where known biological knowledge is stored and well represented, to computationally reconstruct their respective protein–protein interaction maps for functional genomic research.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank two anonymous reviewers for their valuable comments. This work was supported by the National High-Tech Research and Development Program of China (Grant No. 2003AA231030), the Ministry of Education of China (Grant No. 105011) and 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. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
2. Azuaje,F., Wang,H. and Bodenreider,O. (2005) Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*, pp. 9–10.
3. Vezzi,A., Campanaro,S., D'Angelo,M., Simonato,F., Vitulo,N., Lauro,F.M., Cestaro,A., Malacrida,G., Simionati,B., Cannata,N. *et al.* (2005) Life at depth: photobacterium profundum genome sequence and expression analysis. *Science*, **307**, 1459–1461.
4. Hall,N., Pain,A., Berriman,M., Churcher,C., Harris,B., Harris,D., Mungall,K., Bowman,S., Atkin,R., Baker,S. *et al.* (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. *Nature*, **419**, 527–531.
5. Stolc,V., Gauhar,Z., Mason,C., Halasz,G., van Batenburg,M.F., Rifkin,S.A., Hua,S., Herreman,T., Tongprasit,W., Barbano,P.E. *et al.* (2004) A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, **306**, 655–660.
6. Wittkopp,P.J., Haerum,B.K. and Clark,A.G. (2004) Evolutionary changes in *cis* and *trans* gene regulation. *Nature*, **430**, 85–88.
7. Zdobnov,E.M., von Mering,C., Letunic,I., Torrents,D., Suyama,M., Copley,R.R., Christophides,G.K., Thomasova,D., Holt,R.A., Subramanian,G.M. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science*, **298**, 149–159.
8. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
9. The Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
10. Altshuler,D., Brooks,L.D., Chakravarti,A., Collins,F.S., Daly,M.J. and Donnelly,P. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
11. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
12. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
13. 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.
14. Rhodes,D.R., Tomlins,S.A., Varambally,S., Mahavisno,V., Barrette,T., Kalyana-Sundaram,S., Ghosh,D., Pandey,A. and Chinnaiyan,A.M. (2005) Probabilistic model of the human protein–protein interaction network. *Nat. Biotechnol.*, **23**, 951–959.
15. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.
16. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
17. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.
18. Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein–protein interaction data: a probabilistic approach. *Bioinformatics*, **19**, i197–i204.
19. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.
20. Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
21. Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
22. Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
23. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
24. 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.
25. 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.
26. 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.
27. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.
28. 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.
29. 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.
30. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
31. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
32. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale datasets of protein–protein interactions. *Nature*, **417**, 399–403.
33. 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.
34. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.*, **18**, 176–179.
35. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
36. Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.*, **44**, 66–73.

37. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
38. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
39. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
40. Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*., **3**, 199–217.
41. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
42. Qian,J., Dolled-Filhart,M., Lin,J., Yu,H. and Gerstein,M. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol*., **314**, 1053–1066.
43. Jansen,R., Greenbaum,D. and Gerstein,M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res*., **12**, 37–46.
44. Friedman,C., Kra,P., Yu,H., Krauthammer,M. and Rzhetsky,A. (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74–S82.
45. Marcotte,E.M., Xenarios,I. and Eisenberg,D. (2001) Mining literature for protein–protein interactions. *Bioinformatics*, **17**, 359–363.
46. Stapley,B.J. and Benoit,G. (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput*., 529–540.
47. Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol*., **12**, 368–373.
48. 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.
49. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
50. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
51. 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.
52. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Rev. Genet*., **5**, 101–113.
53. Kumar,A. and Snyder,M. (2002) Protein complexes take the bait. *Nature*, **415**, 123–124.
54. Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.
55. Abbott,A. (2002) The society of proteins. *Nature*, **417**, 894–896.
56. Wiwatwattana,N. and Kumar,A. (2005) Organelle DB: a cross-species database of protein localization and function. *Nucleic Acids Res*., **33**, D598–D604.
57. Kumar,A., Agarwal,S., Heyman,J.A., Matson,S., Heidtman,M., Piccirillo,S., Umansky,L., Drawid,A., Jansen,R., Liu,Y. *et al*. (2002) Subcellular localization of the yeast proteome. *Genes Dev*., **16**, 707–719.
58. Ross-Macdonald,P., Coelho,P.S., Roemer,T., Agarwal,S., Kumar,A., Jansen,R., Cheung,K.H., Sheehan,A., Symoniatis,D., Umansky,L. *et al*. (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.
59. Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol*., **20**, 991–997.
60. Bader,J.S., Chaudhuri,A., Rothberg,J.M. and Chant,J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol*., **22**, 78–85.
61. 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.
62. de Lichtenberg,U., Jensen,L.J., Brunak,S. and Bork,P. (2005) Dynamic complex formation during the yeast cell cycle. *Science*, **307**, 724–727.
63. Patil,A. and Nakamura,H. (2005) Filtering high-throughput protein–protein interaction data using a combination of genomic features. *BMC Bioinformatics*, **6**, 100.
64. King,A.D., Przulj,N. and Jurisica,I. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**, 3013–3020.
65. Bu,D., Zhao,Y., Cai,L., Xue,H., Zhu,X., Lu,H., Zhang,J., Sun,S., Ling,L., Zhang,N. *et al*. (2003) Topological structure analysis of the protein–protein interaction network in budding yeast. *Nucleic Acids Res*., **31**, 2443–2450.
66. Robinson,R.C., Turbedsky,K., Kaiser,D.A., Marchand,J.B., Higgs,H.N., Choe,S. and Pollard,T.D. (2001) Crystal structure of Arp2/3 complex. *Science*, **294**, 1679–1684.
67. Akashi,A., Yoshida,Y., Nakagoshi,H., Kuroki,K., Hashimoto,T., Tagawa,K. and Imamoto,F. (1988) Molecular cloning and expression of a gene for a factor which stabilizes formation of inhibitor-mitochondrial ATPase complex from *Saccharomyces cerevisiae*. *J. Biochem. (Tokyo)*, **104**, 526–530.
68. Marres,C.A., de Vries,S. and Grivell,L.A. (1991) Isolation and inactivation of the nuclear gene encoding the rotenone-insensitive internal NADH: ubiquinone oxidoreductase of mitochondria from *Saccharomyces cerevisiae*. *Eur. J. Biochem*., **195**, 857–862.
69. Ahmad,Z. and Sherman,F. (2001) Role of Arg-166 in yeast cytochrome C1. *J. Biol. Chem*., **276**, 18450–18456.
70. Knop,M. and Schiebel,E. (1998) Receptors determine the cellular localization of a gamma-tubulin complex and thereby the site of microtubule formation. *EMBO J*., **17**, 3952–3967.
71. Kaiser,P., Sia,R.A., Bardes,E.G., Lew,D.J. and Reed,S.I. (1998) Cdc34 and the F-box protein Met30 are required for degradation of the Cdk-inhibitory kinase Swe1. *Genes Dev*., **12**, 2587–2597.
72. Krogan,N.J., Kim,M., Ahn,S.H., Zhong,G., Kobor,M.S., Cagney,G., Emili,A., Shilatifard,A., Buratowski,S. and Greenblatt,J.F. (2002) RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell. Biol*., **22**, 6979–6992.
73. Church,C., Chapon,C. and Poyton,R.O. (1996) Cloning and characterization of PET100, a gene required for the assembly of yeast cytochrome *c* oxidase. *J. Biol. Chem*., **271**, 18499–18507.
74. Church,C. and Poyton,R.O. (1998) Neither respiration nor cytochrome *c* oxidase affects mitochondrial morphology in *Saccharomyces cerevisiae*. *J. Exp. Biol*., **201**, 1729–1737.
75. Forsha,D., Church,C., Wazny,P. and Poyton,R.O. (2001) Structure and function of Pet100p, a molecular chaperone required for the assembly of cytochrome *c* oxidase in *Saccharomyces cerevisiae*. *Biochem. Soc. Trans*., **29**, 436–441.
76. Hell,K., Tzagoloff,A., Neupert,W. and Stuart,R.A. (2000) Identification of Cox20p, a novel protein involved in the maturation and assembly of cytochrome oxidase subunit 2. *J. Biol. Chem*., **275**, 4571–4578.
77. Erdemir,T., Bilican,B., Cagatay,T., Goding,C.R. and Yavuzer,U. (2002) *Saccharomyces cerevisiae* C1D is implicated in both non-homologous DNA end joining and homologous recombination. *Mol. Microbiol*., **46**, 947–957.
78. 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.
79. Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**, i38–i46.
80. Ispolatov,I., Yuryev,A., Mazo,I. and Maslov,S. (2005) Binding properties and evolution of homodimers in protein–protein interaction networks. *Nucleic Acids Res*., **33**, 3629–3635.
81. von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A. and Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.