Nucleic Acids Research, 2006, Vol. 34, No. 10 **2943–2952**
doi:10.1093/nar/gkl353

Protein complex compositions predicted by structural similarity

Fred P. Davis[1,2], Hannes Braberg[1,2], Min-Yi Shen[1,2], Ursula Pieper[1,2], Andrej Sali[1,2,*] and M.S. Madhusudhan[1,2,*]

[1]Department of Biopharmaceutical Sciences and [2]Department of Pharmaceutical Chemistry,
California Institute for Quantitative Biomedical Research, University of California, San Francisco,
1700 4th Street, Byers Hall, San Francisco, CA 94143-2552, USA

Received March 15, 2006; Revised April 1, 2006; Accepted April 20, 2006

ABSTRACT

Proteins function through interactions with other molecules. Thus, the network of physical interactions among proteins is of great interest to both experimental and computational biologists. Here we present structure-based predictions of 3387 binary and 1234 higher order protein complexes in *Saccharomyces cerevisiae* involving 924 and 195 proteins, respectively. To generate candidate complexes, comparative models of individual proteins were built and combined together using complexes of known structure as templates. These candidate complexes were then assessed using a statistical potential, derived from binary domain interfaces in PIBASE (http://salilab.org/pibase). The statistical potential discriminated a benchmark set of 100 interface structures from a set of sequence-randomized negative examples with a false positive rate of 3% and a true positive rate of 97%. Moreover, the predicted complexes were also filtered using functional annotation and subcellular localization data. The ability of the method to select the correct binding mode among alternates is demonstrated for three camelid VHH domain—porcine α–amylase interactions. We also highlight the prediction of co-complexed domain superfamilies that are not present in template complexes. Through integration with MODBASE, the application of the method to proteomes that are less well characterized than that of *S.cerevisiae* will contribute to expansion of the structural and functional coverage of protein interaction space. The predicted complexes are deposited in MODBASE (http://salilab.org/modbase).

INTRODUCTION

Recent developments in high-throughput screening have generated large datasets identifying protein complexes. The *Saccharomyces cerevisiae* proteome has been especially well characterized through yeast-two-hybrid (Y2H) (1,2) and tandem affinity purification (TAP) experiments (3–5). Experimentally observed interactions, resulting from both high-throughput and traditional low-throughput methodologies, are deposited in databases such as the Biomolecular Interaction Network Database (BIND) (6) and the Database of Interacting Proteins (DIP) (7).

Concomitant with these experimental advances, a spate of computational techniques to predict protein–protein interactions have also been developed. Several approaches based on protein sequence, structure, function and genomic features have been described (8). In an effort to reduce the prediction errors, several methods integrate multiple types of experimentally determined information and theoretical considerations (9–11).

Structure-based methods have been developed for the prediction of binary protein interactions. InterPreTS (12) uses a statistical potential derived from known hetero-dimer structures and MULTIPROSPECTOR (13) relies on threading to score pairs of proteins that are similar to binary interactions of known structure. In addition to predicting new interactions, structure-based methods can also annotate interactions that have been previously observed experimentally. A recent study used computational methods in conjunction with experimentally determined complex compositions and electron density maps from negative-stain electron cryomicroscopy to generate structural models of yeast complexes (14). In a similar vein, structural knowledge has been used to predict the domains that are most likely to mediate binary protein interactions (15).

Here, we describe predictions of proteins that form complexes in *S.cerevisiae* based on similarity to complexes whose atomic structures have been solved experimentally.

*To whom correspondence should be addressed. Tel: + 1 415 514 4232; Fax: +1 415 514 4231; Email: madhu@salilab.org
*Correspondence may also be addressed to A. Sali. Tel: +1 415 514 4227; Fax: +1 415 514 4231; Email: sali@salilab.org

First, comparative models of conceivable complexes are built and then assessed by a specialized statistical potential. The high-confidence interactions can be additionally filtered by examining orthogonal sources of information including sub-cellular localization and functional annotation.

The current study is unique primarily in its prediction of structural models for higher-order complexes as well as homomeric complexes. Computational methods have been developed to infer higher-order complexes from binary protein interaction networks (16,17), but they do not explicitly use structural knowledge. Previous studies have also focused primarily on the prediction of heterodimers, though homodimerization is biologically prevalent and functionally significant (18). We show that the multiple structure-based assessment steps, from the initial fold assignment, to the interaction prediction, enables our method to achieve a higher coverage, and presumably accuracy, than methods based solely on sequence similarity.

We begin by describing the approach and benchmarking the method. Predictions are then presented for proteins in *S.cerevisiae* and validated against experimentally observed complexes. We highlight the performance of the protocol in the selection of the correct binding mode when multiple template interface structures are available and discuss newly predicted co-complexed superfamilies. Finally, we conclude with a brief discussion of potential applications of the method in light of the ultimate goal offull structural coverage of interaction space.

METHODS

Prediction algorithm

Candidate complexes are first generated, then assessed and finally filtered by orthogonal biological information (Figure 1a).

Candidate complex generation. Pairs of *S. cerevisiae* proteins were identified as potential interaction partners if they were assigned SCOP domains belonging to superfamilies for which an interaction structure exists in PIBASE (Figure 1b) (19). In some superfamilies, such as the ARM superfamily (SCOP a.118.1), the lengths of the member domains vary widely. Because alignments between structures of different lengths are difficult, a threshold was placed on the relative sizes of the target and template domains—the shorter of the two domains must be at least 60% of the length of the longer

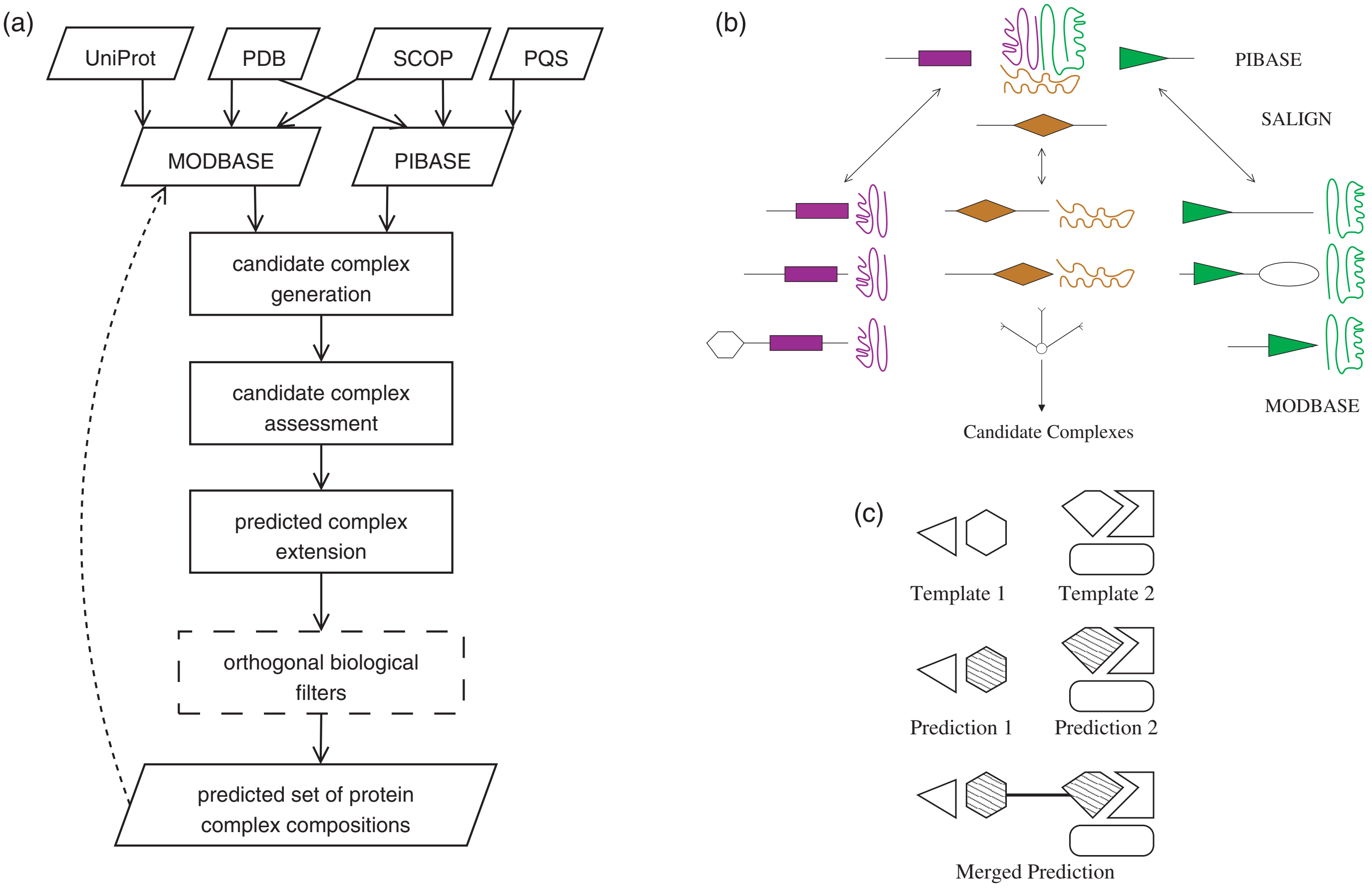

(a)
UniProt
PDB
SCOP
PQS
MODBASE
PIBASE
candidate complex generation
candidate complex assessment
predicted complex extension
orthogonal biological filters
predicted set of protein complex compositions
(b)
PIBASE
SALIGN
Candidate Complexes
MODBASE
(c)
Template 1
Template 2
Prediction 1
Prediction 2
Merged Prediction

Figure 1. Prediction logic overview. (**a**) Prediction flowchart. Groups of protein sequences modeled with SCOP domains observed to form a complex in PIBASE are listed as candidate complexes. These candidate complexes are then assessed by a statistical potential. Interactions that score above a *Z*-score threshold are filtered using sub-cellular localization and functional annotation. The resultant predictions are deposited in MODBASE. (**b**) Candidate complex generation. Comparative models of target domains are structurally aligned to templates of known structure in PIBASE using the SALIGN module of MODELLER. Putative interface residues are identified from the alignment. (**c**) Predicted complexes are merged if they contain different domains of a single target protein.

domain. In addition, the target domains were required to be aligned with the template domains in sufficient number of positions such that the corresponding template residues formed at least 50% of the template interface contacts.

Protein Data Bank (PDB) (20) structures that contained more than two domains were used as templates for the prediction of higher-order complexes with more than two proteins. Target domains that were assessed to interact through the interface modes in a given PDB structure were listed as candidate members of a complex. Each complex was then scored with the worst of the *Z*-scores for the interacting domain pairs it contained, as described below. This was done to provide a conservative estimate of complex quality based on the lowest scoring constituent inerface. Predicted complexes were merged if they contained different domains of a single target protein. In effect, the covalent link between the domains served as a 'bridge' between predicted complexes that were based on different templates (Figure 1c).

Assessment of candidate complexes. Each candidate interaction pair was scored by assessing the agreement between the target sequences and the template interface structure using a statistical potential derived from binary interface structures in PIBASE.

First, residue contacts across the interface were calculated for the template interface and grouped into classes based on the main chain or side chain participation of each residue. Second, the MODBASE models of each candidate interaction partner were structurally aligned against the corresponding domains in the template interface using the SALIGN module of MODELLER (21). Finally, the residue correspondences defined by the alignments were used to score the candidate partner sequences against the template interface contacts using the statistical potential, as described below.

A *Z*-score was calculated to assess the significance of the raw statistical potential score, by consideration of the mean and standard deviation of the statistical potential scores for 1000 sequences where all amino acids in the target domain sequences were shuffled. Sequence randomization has been shown previously to perform comparably with a more physical model involving structural sampling in the context of fold assessment (22).

Orthogonal biological information. Orthogonal biological support for each predicted complex was provided by subcellular localization and gene ontology functional annotation of their components, obtained from the YeastGFP database (23) and SGD (24), respectively. The number of shared localization and function terms were computed for both experimental and predicted complexes. If all pairs of proteins in a complex shared at least one function or localization term, the complex was flagged as co-functioning or co-localized, respectively.

Construction of statistical potentials

A series of statistical potentials was built using the binary domain interfaces in PIBASE extracted from structures at or above 2.5 Å resolution, randomly excluding 100 benchmark interfaces. Twenty-four statistical potentials were built using different values of three parameters: the contacting atom types (main chain–main chain, main chain–side chain, side chain–side chain or all), the relative location of the contacting residues (inter- or intra-domain) and the distance threshold for contact participation (4, 6 or 8 Å):

$$g_{ij} = \frac{\sum_{p=1}^{N} \sum_{c=1}^{\Delta n_{ij}^{(p)}(R_o)} \mathrm{cifa}_{ci,\,cj} n_p}{\sum_{p=1}^{N} n_{ij}^{(p)} \max\left(\mathrm{cifa}_{i,\,j}\right)} \qquad \mathbf{1}$$

$$\mathrm{cifa}_{x,\,y} = \min\left(\frac{\text{interacting atoms}_x}{\text{atoms}_x}, \frac{\text{interacting atoms}_y}{\text{atoms}_y}\right),$$

$$n_{ij}^{(p)} = \begin{cases} n_i^{(p)} n_j^{(p)} & \text{intra-domain potential.} \\ n_i^{(d1)} n_j^{(d2)} + n_i^{(d2)} n_j^{(d1)} & \text{inter-domain potential.} \end{cases}$$

$$w_{ij} = -\ln\left[\frac{g_{ij}}{\frac{1}{400}\sum_{k=1}^{20}\sum_{l=1}^{20} g_{kl}}\right]. \qquad \mathbf{2}$$

Each of the $\Delta n_{ij}^{(p)}(R_o)$ residue pairs of type *i* and *j* in protein *p* that occurred within the distance threshold R_o was weighted by cifa, the minimum of the fraction of total atoms (of the type specified in the potential) in each residue that fell within the distance threshold (Equation 1), and n_p, the number of residues in the protein. This count for each residue type pair was normalized by $n_{ij}^{(p)}$, the total number of possible contacts of that type in each protein, weighted by max(cifa$_{ij}$). In the case of the inter-domain potential, $n_{ij}^{(p)}$ was computed by taking into account the occurrence of each residue type in each domain individually. Finally, the score for each residue type pair was normalized by the sum of the scores observed for all residue type pairs (Equation 2).

Benchmarking of statistical potentials

Performance on the benchmark set of 100 randomly selected interface structures, that were excluded during construction of the potentials, was used to compare the 24 statistical potentials. Of these benchmark interfaces 84 occur between domains from the same family. This is representative of all interfaces in PIBASE, 8877 (15.5%) of which are between domains from different families and 48 257 (84.5%) of which are between domains from the same family. The sequences of the benchmark interfaces were scored against their structures and a *Z*-score was calculated as described above. Receiver–operator curves (ROCs) were built to describe the observed false-positive and true-positive rates at different *Z*-score thresholds. ROCs were then integrated to calculate the area under the curve (AUC). The AUC represents the probability that a classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance, with 0.5 corresponding to a random prediction, and 1 to a perfect classifier (25) (www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf)

To investigate the effect of variation in the benchmark set on each of the ROCs, 20 jack-knife trials were performed where 20 randomly selected structures were removed and the ROCs recomputed using the remaining 80 structures.

Standard deviations of the areas under the resulting ROCs were then calculated.

Validation of complex prediction

The predicted interactions were validated in two ways. First, the predicted *S.cerevisiae* complexes were compared with the experimentally determined complexes in the BIND database (6) and those reported recently by Gavin *et al*. (3) referred to as Cellzome. The binary interactions were compared by counting the overlap of the predictions with the interactions in the BIND and Cellzome sets. The Cellzome set consisted of pairs of proteins that were deemed highly reliable in forming partnerships based on their computed 'socio-affinity' score (3).

Second, the higher order complexes were compared between the predicted and experimental sets by counting how many of the predicted complexes were equivalent to, or were subcomplexes of, experimentally determined complexes. Since the predictions are based on known structures, the sizes of the predicted complexes are far smaller than those obtained by biochemical methods such as tandem affinity purification methods. For this reason, we elected not to use a metric that explicitly penalizes size differences [e.g. the metric defined in Ref. (16)].

Binding mode selection

The ability of the potential to select the proper binding mode when multiple template interfaces of different orientation are available was assessed. The test cases used were the structures of camelid VHH domains AMB7, AMD10 and AMD9 bound to porcine pancreatic α-amylase (PPA) (PDB codes 1kxt, 1kxv and 1kxq, respectively). All three modes were evaluated for each VHH–PPA complex using the interface statistical potential.

Data sources

The prediction algorithm uses three types of data: (i) target protein sequences among which complexes are to be predicted, (ii) structures of protein complexes to be used as templates, and (iii) a list of the locations and types of structural domains in the target and template proteins (Figure 1a).

Target proteins. *S.cerevisiae* protein sequences were obtained from MODBASE, a relational database of annotated comparative protein structure models for all available protein sequences matched to at least one known protein structure (26). The models were calculated by MODPIPE (27), an automated modeling pipeline that relies on MODELLER for fold assignment, sequence–structure alignment, model building and model assessment (21). *S.cerevisiae* proteins (6600) were processed, resulting in 9464 models for 3440 sequences. A total of 2659 sequences had at least one reliable model (5387 reliable models in total). A model is considered reliable when the model score, derived from statistical potentials, is higher than a cutoff of 0.7 (22). A reliable model has >95% probability of having at least 30% of Cα atoms within 3.5 Å of their correct positions. A total of 3376 sequences had at least one reliable fold assignment (8935 reliable folds in total). A fold assignment is considered reliable when the model is based on a PSI-BLAST match to a template with an *E*-value <0.0001.

Structural domain annotation. The domain definitions for PDB structures were obtained from the SCOP database (ver 1.69) that classifies each domain using a four level hierarchy, class, fold, superfamily and family (28). The location and types of domains in the target protein sequences were then predicted using the SCOP annotation of their MODBASE templates, as follows. Domain boundaries were first assigned based on the MODBASE alignment of each target protein to its structural template. Each target domain was required to have at least 70% of the residues in its template domain to receive the domain assignment. Next, if the target domain had >30% sequence identity to the template domain and the MODBASE structural model was assessed to be reliable, the target domain received the template's SCOP classification at the family level. If the sequence identity was <30% and a reliable model was built or if the sequence identity was >30% but MODBASE deemed only a reliable fold assignment, the superfamily was assigned. The remaining domains received the template domain's SCOP classification at the fold level, and were not used in the interaction prediction.

For those target proteins for which multiple models were available in MODBASE, a tiling procedure combined the domain assignments for each model into a non-overlapping set of domain boundaries that maximized the coverage length and classification detail in the SCOP hierarchy.

Template complexes. Structures of template complexes were retrieved from PIBASE, a comprehensive relational database of structurally defined protein interfaces (19). It currently includes 209 961 structures of interactions between 2613 SCOP domain families. The ASTEROIDS component of the SCOP ASTRAL compendium was used to cluster the interfaces, reducing the computational expense of the predictions (29). The ASTEROIDS alignments, available for SCOP classes a–g, were used together with the interface contacts stored in PIBASE to cluster all interface structures that shared pairs of SCOP families. When at least 75% of the pairwise residue contacts in one interface also occurred between residues that were aligned in another interface, the two interfaces were merged into a single cluster. The clustering reduced the 79 428 domain interfaces between pairs of domains in the SCOP classes a–g to 21 791 representative interfaces. These interfaces were filtered using a threshold of at least 1000 interatomic contacts resulting in a set of interfaces of significant size. Thresholds similar to this, which roughly corresponds to a buried surface area of 400 A^2, have been used previously to filter crystallographic artifacts from biologically relevant interfaces (30). The final set of template binary interfaces contained 5275 structures, including both intermolecular and intramolecular interfaces.

Technology

The prediction system was implemented as a Perl module and an integrated set of Perl scripts, except for the inter-atomic contacts calculator written in ANSI C (19). The SALIGN module of MODELLER (21) was used to generate model template alignments. The Perl DBI interface was used to access the MODBASE and PIBASE MySQL databases (http://www.mysql.com). The calculations were done in a parallel fashion on 50 3.0 GHz Pentium IV processors, taking 20 h for the yeast genome. The predictions are accessible via the MODBASE web interface (http://salilab.org/modbase).

RESULTS

Benchmark

The statistical potentials were tested using the benchmark set of 100 complexes, and their performance compared using (ROCs) (Methods). The highest power of discriminating between the native and non-native interfaces was achieved by the statistical potential built from side chain–side chain contacts across the interfaces at a threshold of 8 Å, corresponding to the extent of the first residue shell (Figure 2). The ROC for this potential had an area under the curve (AUC) of 0.993, and at the optimal *Z*-score threshold of −1.7 had true positive and false positive rates of 97 and 3%, respectively. Clear performance trends were observed for the parameters sampled in the potential construction. The general trend of increased performance at the 8 Å threshold over the lower thresholds is likely due to a more complete description of interactions within the first residue shell. The inter-domain potential always performed better than the corresponding intra-domain potential, when all other parameters were equivalent (data not shown). The side chain–side chain (SS) potential performed better than the corresponding main chain–side chain (MS) potential, which in turn performed better than the corresponding main chain–main chain (MM) potential. At 6 and 8 Å, the all atom-type potential performed better than only the MM potential. At 4 Å, the all atom-type potential performed better than both MS and MM potentials. The range of performances, generatedby varying the other parameters (i.e. atom type, inter- or intra-domain), was widest at the 4 Å distance threshold and least at 8 Å.

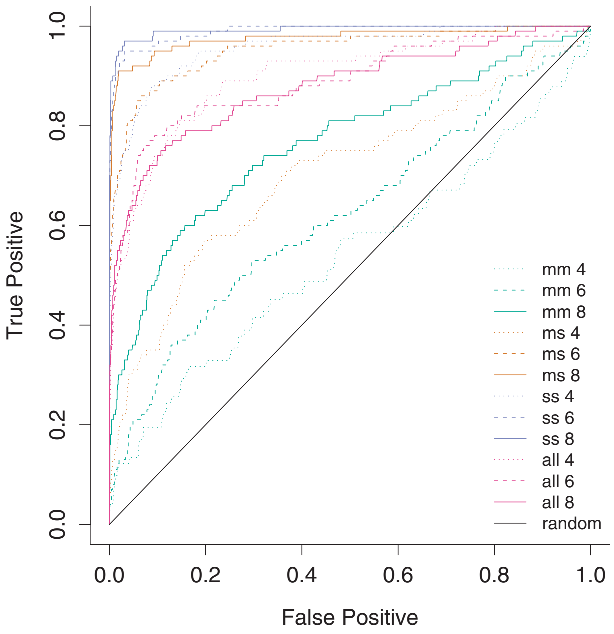


Figure 2. Assessment of statistical potentials. (ROCs) are shown for the inter-domain potential performance on the benchmark set of complexes.

Jack-knife trials were performed to determine the effect of variation in the benchmark set on the ROCs (Methods). The AUC of the jack-knifed ROCs exhibited narrow distributions, with the lowest standard deviation (0.002) achieved by the inter-domain SS potential at 8 Å, and the highest (0.02) achieved by the intra-domain MM potential at 4 Å. This suggests the ROC analysis is robust to variations in the benchmark set.

Here, the potentials were assessed using a benchmark set of native interface structures. In the predictive setting, the absolute performance of each potential will likely be diminished due to errors in the comparative models. However, the relative performance of the different formulations, as captured by the ROCs, remains a valid guide for selection of the potential to use in the predictions.

Predictions

The best statistical potential, as determined above, was then used to assess candidate interactions between *S.cerevisiae* proteins. A total of 12 867 binary interactions that scored at or below a *Z*-score threshold of −1.7 were predicted between 1390 *S.cerevisiae* proteins (Figure 3a). Next, the co-function and co-localization filters were separately applied, reducing the original 12 867 interactions to 6808 and 4606, respectively. The combined co-localization and co-function filter resulted in 3387 predictions. A total of 12 702 higher-order complexes were also predicted at a *Z*-score threshold of −1.7 between 589 proteins. Similar to the binary predictions, the orthogonal filters reduced this number to 1234 complexes between 195 proteins.

The predictions spanned the entire spectrum of target–template sequence similarity (Figure 3b). This distribution reflects both the comparative modeling procedure used to build models of the individual proteins and the procedure used to identify potential interaction templates. The mean target–template sequence identity of the reliable models built for *S.cerevisiae* proteins is 31%. Domains from different families within the same superfamily, the SCOP level used to identify potential interaction templates, often share <30% sequence identity. Both of these factors influence the distribution of target–template identities observed for the predicted interactions.

The fractions of predicted binary interactions that passed the co-function (53%), co-localization (36%), and both co-function and co-localization (26%) filters were similar to the fractions for BIND interactions (39, 34 and 22%, respectively). The Cellzome set more readily passed these filters (65, 60 and 46%, respectively).

Validation

The predictions were then compared with known experimental interactions, as deposited in the BIND database. Of the 3387 predicted binary interactions that passed the combined co-localization and co-function filter 270 overlapped with known binary interactions. Of the 1234 predicted higher-order complexes 8 were also found as subcomplexes of experimental complexes.

(a)

	Protein Interactions	Proteins	Domain Interfaces	Domains
Input				
MODBASE models	-	3,440	-	5,219
Template Complexes	-	-	5,275	9,314
Binary Interactions				
Z-score ≤ −1.7	12,867 (red5.1%)	1,390	13,773	1,727
Z + Co-Function	6,808 (red9.6%)	1,152	7,364	1,389
Z + Co-Localization	4,606 (red14.1%)	1,021	5,030	1,255
Z + Co-Loc + Co-Func	3,387 (red19.2%)	924	3,738	1,112
Higher-Order Complexes				
Z-score ≤ −1.7	12,702	589		
Z + Co-Function	3,544	332		
Z + Co-Localization	2,189	280		
Z + Co-Loc + Co-Func	1,234	195		

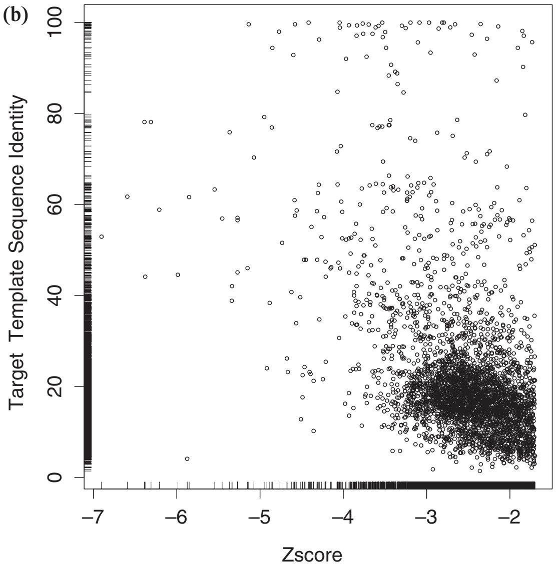
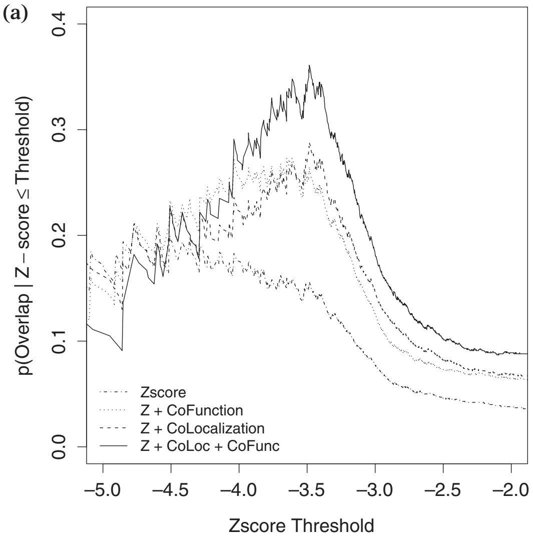


Figure 3. *S.cerevisiae* predictions. (**a**) Predictions of binary and higher-order complexes filtered by subcellular localization and annotated function. The homomeric fraction of interactions is listed in parenthesis. (**b**) Average sequence identity of predicted interaction partners to template interacting domains versus Z-score. The predictions shown were scored with *Z*-score ≤−1.7, and passed the combined co-localization and co-function filter.



(b)

		Experimental Overlap		
	Predicted	All	BIND	Cellzome
Binary Interactions				
experimental		19,424	13,191	6,942
Z-score ≤ −1.7	12,867	409	324	151
Z + Co-Function	6,808	390	311	145
Z + Co-Localization	4,606	278	220	102
Z + Co-Loc + Co-Func	3,387	270	217	97
Higher-Order complexes				
experimental		783	296	491
Z-score	12,702	66	54	35
Z + Co-Function	3,544	51	45	28
Z + Co-Localization	2,189	14	7	10
Z + Co-Loc + Co-Func	1,234	8	4	7

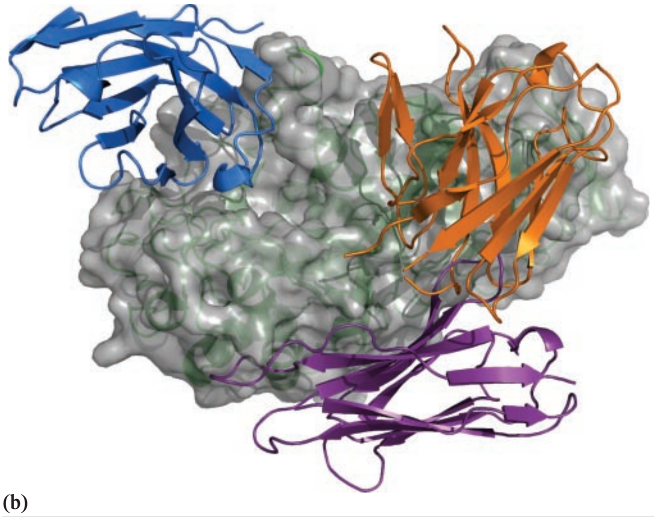
Figure 4. Experimental overlap of *S.cerevisiae* predictions. (**a**) The probability of finding an experimentally observed interaction in the predicted set, as a function of the statistical potential *Z*-score. The unfiltered predictions are represented by dotted-dashed, the co-function filtered by dotted, the co-localization by dashed, and the combined co-localization and co-function filtered set by solid lines. The curves are only shown to a *Z*-score threshold of −5.0, because of the sparseness of predictions below this level. (**b**) Experimental overlap of the binary and higher-order predictions filtered by sub-cellular localization and annotated function.

The enrichment of the unfiltered predictions with known binary interactions begins to plateau at 0.15 around a *Z*-score threshold of −3.5, with an enrichment value of 0.03 at the *Z*-score of −1.7 (Figure 4a). The predictions that passed the separate localization and function filters both reached a peak of 0.28 at a *Z*-score of −3.5. Both filters produced enrichment values of 0.06 at the *Z*-score threshold of −1.7. The enrichment of the predictions that passed the combined co-localization and co-function filter exhibited a higher peak of 0.36 at the *Z*-score of −3.5. At the *Z*-score threshold of −1.7, the combined filter produced an enrichment of 0.08, a >3-fold increase compared with the unfiltered predictions.

Comparison with other computational methods

The performance of the method in predicting binary interactions is comparable with similar structure-based methods that

(a)



(b)

	AMB7 mode	AMD10 mode	AMD9 mode	K_d [nM]
AMB7	*-3.27 (-2.27)*	-1.19 (14.02)	-2.65 (5.00)	235
AMD10	-1.39 (12.61)	*-3.40 (-4.84)*	-2.36 (6.73)	25
AMD9	-2.13 (4.94)	-0.97 (15.78)	*-3.60 (-9.75)*	3.5

Figure 5. (**a** and **b**) Selection among alternate binding modes. Camelid VHH domains AMB7 (orange), AMD10 (magenta) and AMD9 (blue) bind to porcine pancreatic α-amylase (PPA, gray surface) through three distinct binding modes (PDB codes 1kxt, 1kxv, and 1kxq, respectively). All three modes were evaluated for each VHH–PPA complex using the interface statistical potential. The *Z*-scores are presented along with the raw score in parenthesis. Dissociation constants measured by total internal reflectance (IAsys) were obtained from literature (33). Image created by PyMOL (Delano Scientific, 2002).

have been applied previously to *S.cerevisiae* on a genomic scale. Here, an overlap of 270 binary interactions is observed between the set of 3387 (8%) predictions and 19 424 (1.4%) experimentally observed binary interactions. Of 7321 (5%) interactions predicted by threading 374 occurred in a set of 78 930 (0.4%) experimentally determined yeast interactions (31). An overlap of 59 predicted interactions with an experimental set of 2590 (2.3%) interactions was obtained by interface model assessment (12).

To compare it directly with a method that does not use structural assessment, PSI-BLAST (32) was used to predict binary interactions by detecting similarities between *S.cerevisiae* proteins and the template complexes. An overlap of 929 binary interactions was observed between the set of 36 790 (2.5%) predictions and the 19 424 (4.8%) experimentally observed binary interactions.

Alternate binding modes

The ability of the algorithm to correctly select the native binding mode when alternate templates are available was tested. The native binding mode was correctly selected for all three VHH domains interacting with porcine pancreatic α-amylase (Figure 5). In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank-order as the affinity measured experimentally by total internal reflectance (33).

	Superfamily pairs	BIND or Cellzome	BIND	Cellzome
BIND or Cellzome	13,586	13,586	3,997	11,594
PDB direct	671	181	131	159
PDB co-complexed	1,555	420	143	393
Predicted co-complexed	100	43	24	35

Figure 6. Co-complexed domain superfamilies. The pairs of co-complexed superfamilies observed in the BIND and Cellzome complexes are compared with the direct interactions in the PDB, co-complexed pairs in the PDB and the predicted co-complexed pairs resulting from the complex extension procedure.

Co-complexed domains

An extension process merged predicted complexes containing different domains of a single target protein (Figure 1c). This process predicted 279 pairs of co-complexed SCOP domain families that were not present in the structures of template complexes. The comparison with experimental complexes was done at the superfamily level, as many of the domains in the experimental complexes were assigned domains that were classified only to this level in the SCOP hierarchy (Figure 6).

DISCUSSION

We presented a method to predict protein complex compositions by generating comparative models of candidate complexes based on sequence similarity to structurally known complexes followed by model assessment (Figure 1). We applied the method to the *S.cerevisiae* proteome (Figure 3) and compared the predicted complexes with experimental data (Figures 4 and 6). We further tested the method by distinguishing between multiple template binding modes (Figure 5). We now discuss the observed performance and describe the limitations of the algorithm. We close by discussing the information gained in the present study and its applications to increasing structural description of protein interactions.

Accuracy

Because a large set of true negative interactions is not available, only the positives, or predicted interactions, can be compared between experiment and predictions. This limitation restricts the validation of the predictions because if the *Z*-score threshold is loosened, maximal overlap can be achieved at the expense of the false positive rate. However, the false positive rate cannot be counted with certainty, as false positives cannot be distinguished from false negatives in the experimental datasets, which can be quite high (34). Similar validation problems are encountered when testing protein ligand docking algorithms. Here, a measure related to the enrichment factor used in protein ligand docking was applied (Figure 4a).

The overlap observed between the predicted and experimentally observed complexes is comparable with that between different experimental procedures (34). Of the 3387 predicted binary interactions 270 and of the 1234 predicted higher-order complexes 8 were present in the BIND or Cellzome datasets (Figure 4).

This overlap is a result of several factors. First, by construction our method is restricted to protein interactions for which structural templates exist. For this reason, our method is also biased towards complexes that are stable enough to be amenable to structure determination, whereas the Y2H method that populates most of the high-throughput entries in BIND, is biased towards transient interactions (34). Second, many PDB entries do not contain complete domains for both partners (e.g. SH3 domain–peptide complexes) and were thus not considered as templates in the current prediction protocol. Finally, the challenge faced in predicting binary interactions increases combinatorially for higher-order complexes.

Errors in the predicted interactions are also a result of errors that may arise in each stage of the comparative modeling procedure, including fold assignment, alignment and structure modeling. Comparative modeling errors vary in type and magnitude with the sequence identity between the template and target proteins (35). At very low sequence identities, the fold type of the target sequence may be assigned erroneously. When the proper fold has been assigned, misalignments may still occur due to gaps and insertions in the target sequence. Given the correct alignment, main chain distortions may still occur due to differences in the target and template backbone structure. At the finest resolution, side chains may suffer from errors in packing. These comparative modeling errors contribute to both false positives and false negatives in the predicted interactions.

The use of sub-cellular localization data and functional annotation as filters for the predictions increased their overlap with experimental complexes, as compared with the unfiltered predictions. This finding is in agreement with previous observations that combining multiple sources of information improves the accuracy of function annotation as well as interaction prediction (9–11). Our method easily allows for the use of additional biological filters when other types of data are available, such as synthetic gene lethality (36), co-expression (37), and so on. This incremental addition of orthogonal information is also necessary to more accurately represent the conditions in the cellular milieu, where the propensity of two protein structures to interact is not limited only by the physical chemistry of the interaction, but also by higher levels of biological regulation, including compartmentalization, expression, degradation, abundance and so on. Depending on the application, the user may decide to apply different biological filters.

Importance of structure

The majority (98.6%) of the filtered binary interactions as well as the subset that overlapped with experimentally observed interactions (86.9%) were based on templates sharing <80% sequence identity, a threshold established previously for reliable transfer of a known interaction to a putative interaction between homologous proteins (Figure 3b) (38). This distribution highlights the advantage garnered by the use of structure and the importance of a structure-based assessment.

One such example is the experimentally observed interaction between LSM2 and LSM7 that was predicted here based on structural similarity to the 14mer complex of SmAP3, an Sm-like protein from the archae *Pyrobaculum aerophilum* (PDB 1m5q). The sequence identities of LSM2 and LSM7 to SmAP3 are 23 and 2.4%, respectively. While interface templates with higher sequence identities were available (highest identities of 20.7% for LSM2 and 32.1% for LSM7 to chains G and A of PDB 1jbm, respectively), the1m5q-based model was scored most favorably by the statistical potential. Another example of a known interaction predicted using a distantly related template interaction is that between the delta (GCD2) and beta (GCD7) subunits of the translation initiation factor eIF2B, predicted based on similarity to the structure of Ypr118w, a methylthioribose-1-phosphate isomerase related to regulatory eIF2B subunits. The prediction was made based on sequence similarities of 16 and 15%, respectively.

For comparison, a naïve search for putative interaction partners was performed by using PSI-BLAST to detect similarities between yeast proteins and the template complexes. As expected, this approach, which is equivalent to the current method performed without the structural assessment, predicted more binary interactions that have been observed previously by experiment (929) than the structure-based method (270). However, the naïve approach likely suffers from a higher false positive rate, as can be observed in the lower enrichment of its predictions with experimentally observed interactions (2.5%) than the structure-based method (8%) (Methods).

Alternative binding modes

The ability of the algorithm to choose the correct binding mode when multiple templates are available was illustrated by evaluation of three alternative binding modes that have been structurally characterized between porcine pancreatic α-amylase and camelid VHH domains (Figure 5). The algorithm successfully chose the native binding mode for all three VHH domains. In addition, the statistical potential scores that were computed for the native binding modes exhibit the same rank order as the affinity of the interactions measured by total internal reflectance (33).

However, this example is also cautionary in that each VHH domain had one non-native mode that scored below the optimal *Z*-score threshold, though only the native modes produced negative raw scores (Results). In a large-scale predictive setting, if the native binding mode was not available as a template, the VHH domain would have been predicted to interact with PPA, but through an incorrect binding mode. This example illustrates a connection between the observed performance and the underlying scoring scheme. However, a systematic analysis of alternative binding modes in protein interactions, and the ability of our method to distinguish them, remains a useful goal for the future.

Network specificities

A more difficult test of the method is the prediction of specificities within interaction networks between homologous proteins. To address this problem, the method was applied to predict the specificities within the epidermal growth factor receptor (EGFR) and tumor necrosis factor β (TNFβ) networks of ligand receptor interactions (data not shown). In both networks the method failed to recapitulate known binding preferences. Specifically, the rank order of the *Z*-scores for the assessed pairs did not correlate with known binding preferences.

This error was not surprising. The randomization scheme employed in the *Z*-score assessment of the raw statistical potential score simulated alternative binding modes. In contrast, it was not designed or tested to determine specificities. This task is difficult as large training datasets of this type are not available.

Rather than predicting specificities, the method presented here is applicable as a first pass for genome-wide predictions of protein complexes. The resulting predictions are then suitable for a follow up with more accurate computational methods, which on their own are not feasible on a large scale.

Extension of known co-complexed domain superfamilies

Large protein complexes present unique challenges to structural characterization. Direct physical interactions have been experimentally observed between domains from 671 pairs of different SCOP superfamilies (excluding homo-family interactions). Domains from 1555 pairs of different superfamilies have been observed to co-complex in the same PDB entry. Of these pairs 420 have also been observed in biochemical complexes. Through an extension process that merged predicted complexes containing different domains of a single target protein, an additional 100 pairs of superfamilies were predicted to be co-complexed (Figure 1c and 6). Of these newly predicted pairs 43 were also found in the experimental complexes. This extension procedure will be especially informative when applied to proteins from higher organisms with greater domain architecture complexity than *S.cerevisiae* (39).

Future directions

We presented a tool for the prediction and assessment of the composition and structure of protein complexes. The results suggest that the algorithm may in practice be useful in conjunction with additional biological data, such as protein localization and functional annotation. Through its integration with MODBASE, the method is applicable, in an automated fashion, to all genomes with sequences that are amenable to comparative protein structure modeling. The method will be especially informative for proteomes of species that have not been characterized to the extent of *S.cerevisiae*, either because the genomes have only been sequenced recently or because the organisms are difficult to analyze experimentally.

In addition to proposing new protein complexes that have not been observed previously, the present study also enables a more rigorous, structure-based, analysis of experimental protein interaction data. For instance, the system could be used to distinguish complexes from temporally distinct interactions by assessing whether the interactions are sterically compatible or exclusive (40). The predictions may also prove useful in guiding experiments that aim to probe the interactions, such as various site-directed mutagenesis and interaction design studies.

Comparative protein structure modeling is increasingly used to help bridge the resolution gap between electron cryo-microscopy (cryo-EM) density maps and atomic protein structures (41). Fitting of protein and protein domain models into density maps of large assemblies is already common, but depending on the resolution, the information encoded in the map is often insufficient for an unambiguous determination of the positions and orientations of the individual proteins (42). Models of the complexes predicted here may provide additional restraints for a more accurate fitting of proteins into large complexes studied by cryo-EM and electron cryo-tomography (14,43).

As the number and size of experimentally determined structures of protein complexes increase, the number of complexes that can be predicted and modeled using these structures as templates increases correspondingly, expanding the structural coverage of protein interaction space (44). In combination with other computational methods, the presented method will allow biologists to harness interaction information that has been experimentally determined for similar systems to inform their hypotheses or experiments.

ACKNOWLEDGEMENTS

We would like to thank members of the Sali laboratory for valuable comments and suggestions, in particular Maya Topf, Eswar Narayanan and Frank Alber. We acknowledge funding by the Sandler Family Supporting Foundation, NSF (EIA-0325004), NIH (AI035707, U54 RR022220), as well as computer hardware gifts from Sun, Intel and IBM. F.P.D. acknowledges support from a Howard Hughes Medical Institute predoctoral fellowship. Funding to pay the Open Access publication charges for this article was provided by NIH U54 RR022220.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S., Dumpelfeld,B. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature.*, **440**, 631–636.
4. 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.
5. 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.

6. Bader,G.D., Betel,D. and Hogue,C.W. (2003) Bind: the biomolecular interaction network database. *Nucleic Acids Res.*, **31**, 248–250.
7. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
8. Salwinski,L. and Eisenberg,D. (2003) Computational methods of analysis of protein–protein interactions. *Curr. Opin. Struct. Biol.*, **13**, 377–382.
9. 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.
10. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
11. 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.
12. Aloy,P. and Russell,R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
13. Lu,L., Lu,H. and Skolnick,J. (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins*, **49**, 350–364.
14. Aloy,P., Bottcher,B., Ceulemans,H., Leutwein,C., Mellwig,C., Fischer,S., Garin,A.C., Bork,P., Superti-Furga,G., Serrano,L. *et al.* (2004) Structure-based assembly of protein complexes in yeast. *Science*, **303**, 2026–2029.
15. 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.
16. Bader,G.D. and Hogue,C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
17. Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
18. Marianayagam,N.J., Sunde,M. and Matthews,J.M. (2004) The power of two: protein dimerization in biology. *Trends Biochem Sci.*, **29**, 618–625.
19. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
20. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
21. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
22. Melo,F., Sanchez,R. and Sali,A. (2002) Statistical potentials for fold assessment. *Protein Sci.*, **11**, 430–448.
23. Ghaemmaghami,S., Huh,W.K., Bower,K., Howson,R.W., Belle,A., Dephoure,N., O'Shea,E.K. and Weissman,J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
24. Dwight,S.S., Harris,M.A., Dolinski,K., Ball,C.A., Binkley,G., Christie,K.R., Fisk,D.G., Issel-Tarver,L., Schroeder,M., Sherlock,G. *et al.* (2002) *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the gene ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.
25. Fawcett,T. (2003) ROC graphs: notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs, Palo Alto, CA, USA.
26. Pieper,U., Eswar,N., Davis,F.P., Braberg,H., Madhusudhan,M.S., Rossi,A., Marti-Renom,M., Karchin,R., Webb,B.M., Eranian,D. *et al.* (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **34**, D291–D295.
27. Eswar,N., John,B., Mirkovic,N., Fiser,A., Ilyin,V.A., Pieper,U., Stuart,A.C., Marti-Renom,M.A., Madhusudhan,M.S., Yerkovich,B. *et al.* (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.*, **31**, 3375–3380.
28. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
29. Chandonia,J.M., Hon,G., Walker,N.S., Conte,L.L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.
30. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
31. Lu,L., Arakaki,A.K., Lu,H. and Skolnick,J. (2003) Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.*, **13**, 1146–1154.
32. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
33. Lauwereys,M., Ghahroudi,M.A., Desmyter,A., Kinne,J., Holzer,W., Genst,E.D., Wyns,L. and Muyldermans,S. (1998) Potent enzyme inhibitors derived from dromedary heavy-chain antibodies. *EMBO J.*, **17**, 3512–3520.
34. 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.
35. Madhusudhan,M.S., Marti-Renom,M.A., Eswar,N., John,B., Pieper,U., Karchin,R., Shen,M.Y. and Sali,A. (2005) Comparative Protein Structure Modeling. In Walker,J.M. (ed.), *The Proteomics Protocols Handbook.*. Humana Press Inc., Totowa, NJ, pp. 831–860.
36. Tong,A.H., Evangelista,M., Parsons,A.B., Xu,H., Bader,G.D., Page,N., Robinson,M., Raghibizadeh,S., Hogue,C.W., Bussey,H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.
37. Tirosh,I. and Barkai,N. (2005) Computational verification of protein-protein interactions by orthologous co-expression. *BMC Bioinformatics*, **6**, 40.
38. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. and Gerstein,M. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.*, **14**, 1107–1118.
39. Bornberg-Bauer,E., Beaussart,F., Kummerfeld,S.K., Teichmann,S.A. and Weiner,J.,III (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol. Life Sci.*, **62**, 435–445.
40. Han,J.D., Bertin,N., Hao,T., Goldberg,D.S., Berriz,G.F., Zhang,L.V., Dupuy,D., Walhout,A.J., Cusick,M.E., Roth,F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.
41. Topf,M. and Sali,A. (2005) Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.*, **15**, 578–585.
42. Fabiola,F. and Chapman,M.S. (2005) Fitting of high-resolution structures into electron microscopy reconstruction images. *Structure*, **13**, 389–400.
43. Sali,A., Glaeser,R., Earnest,T. and Baumeister,W. (2003) From words to literature in structural proteomics. *Nature*, **422**, 216–225.
44. Aloy,P. and Russell,R.B. (2004) Ten thousand interactions for the molecular biologist. *Nat. Biotechnol.*, **22**, 1317–1321.