MULTIPROSPECTOR: An Algorithm for the Prediction of Protein–Protein Interactions by Multimeric Threading

Long Lu,[1,2] **Hui Lu,**[1] **and Jeffrey Skolnick**[1*]
[1]*Laboratory of Computational Genomics, Donald Danforth Plant Science Center, St. Louis, Missouri*
[2]*Department of Biochemistry and Molecular Biophysics, School of Medicine, Washington University, St. Louis, Missouri*

ABSTRACT **In this postgenomic era, the ability to identify protein–protein interactions on a genomic scale is very important to assist in the assignment of physiological function. Because of the increasing number of solved structures involving protein complexes, the time is ripe to extend threading to the prediction of quaternary structure. In this spirit, a multimeric threading approach has been developed. The approach is comprised of two phases. In the first phase, traditional threading on a single chain is applied to generate a set of potential structures for the query sequences. In particular, we use our recently developed threading algorithm, PROSPECTOR. Then, for those proteins whose template structures are part of a known complex, we rethread on both partners in the complex and now include a protein–protein interfacial energy. To perform this analysis, a database of multimeric protein structures has been constructed, the necessary interfacial pairwise potentials have been derived, and a set of empirical indicators to identify true multimers based on the threading Z-score and the magnitude of the interfacial energy have been established. The algorithm has been tested on a benchmark set comprised of 40 homodimers, 15 heterodimers, and 69 monomers that were scanned against a protein library of 2478 structures that comprise a representative set of structures in the Protein Data Bank. Of these, the method correctly recognized and assigned 36 homodimers, 15 heterodimers, and 65 monomers. This protocol was applied to identify partners and assign quaternary structures of proteins found in the yeast database of interacting proteins. Our multimeric threading algorithm correctly predicts 144 interacting proteins, compared to the 56 (26) cases assigned by PSI-BLAST using a (less) permissive E-value of 1 (0.01). Next, all possible pairs of yeast proteins have been examined. Predictions (n = 2865) of protein–protein interactions are made; 1138 of these 2865 interactions have counterparts in the Database of Interacting Proteins. In contrast, PSI-BLAST made 1781 predictions, and 1215 have counterparts in DIP. An estimation of the false-negative rate for yeast-predicted interactions has also been provided. Thus, a promising approach to help assist in the assignment of protein–protein interactions on a genomic scale has been developed. Proteins 2002;49:350–364.**

Key words: protein–protein interactions; threading; interfacial potentials; protein dimers; genomic scale predictions

INTRODUCTION

Protein–protein interactions are fundamental to cellular function and are associated with processes such as enzymatic activity, immunological recognition, DNA repair and replication, and cell signaling.[1] Often a given protein's function can be inferred from the nature of the proteins with which it interacts. Because of the enhancement in biological information that can be provided by knowledge of protein–protein interactions, this problem has been extensively studied.[2–5] In particular, experimental techniques that are designed to study protein–protein interactions have become more mature and accurate. Qualitative methods designed to determine whether two proteins interact include the yeast two-hybrid screen,[6] immunoprecipitation,[7] and gel-filtration chromatography.[8] Protein–protein interactions can also be quantitatively measured by biophysical methods such as analytical ultracentrifugation,[9] calorimetry,[10] and optical spectroscopy.[11] Ultimately, a protein complex could be crystallized and its quaternary structure determined. Alternatively, the presence and identity of interacting residues could be determined from recently developed NMR techniques.[12,13] However, these experimental techniques are very labor intensive. This has spurred the development of computational algorithms to automatically predict protein–protein interactions.[14] In this spirit, the development of a threading based approach to quaternary structure prediction is described and validated in this article.

Over the past decade, various quaternary structure prediction approaches have been developed. One method focuses on locating interaction sites, without knowing the identity of the specific binding partners.[15] This method uses properties related to the topology of the interface, the

Grant sponsor: National Institutes of Health Division of General Medical Sciences; Grant number: GM-48835.

*Correspondence to: Dr. Jeffrey Skolnick, Donald Danforth Plant Science Center, 975 N. Warson Rd., St. Louis, MO 63132. E-mail: skolnick@danforthcenter.org

Received 5 April 2002; Accepted 25 June 2002

solvent-accessible surface area (ASA), and hydrophobicity.[16] Another structure-based prediction technique, docking, requires the knowledge of the tertiary structure of both partners before predicting the quaternary structure.[17–19] To make a realistic prediction, the structure of the isolated chains should be in the unbound form (i.e., that which each protein adopts in the absence of the other partner). There are two types of docking approaches: geometry-based[20] and energy-based.[21] Several docking programs that use one or both types, such as FTDOCK[22] and GRAMM,[23] are publicly available.

Both binding site identification and docking approaches have other limitations. With respect to binding, protein–protein interactions are quite diverse with no general rules describing how proteins bind.[24] Because quite a large portion of mutual protein binding involves induced fit with the shape of the surface that will involve interactions changing on binding,[25] quaternary structure prediction based on the structure of the isolated, noninteracting monomers could be significantly in error. Furthermore, both the pure geometry-based and the energetic based docking methods are computationally expensive. Often docking algorithms cannot assess which proteins interact and which do not. Because it usually takes hours to predict the interacting sites for a pair of potentially interacting proteins, at present, it is impractical to use docking to predict protein–protein interactions for a large number of proteins. Docking is also limited to those proteins whose structures have already been solved; again, there is the problem of induced fit.[25]

With the successful genome sequencing efforts,[26] some pure sequence-based approaches have begun to predict protein–protein interactions.[27] A domain fusion analysis has been proposed for inferring protein interactions from genome sequences based on the observation that some pairs of interacting proteins have homologues in other organisms that are fused into a single protein chain.[27] Because this method is sequence based, the interaction sites cannot be directly identified (i.e., which pairs of residues form the protein–protein interface). Furthermore, this method fails to indicate the significance of each prediction. Nevertheless, this approach is important in that it was one of the first attempts to predict protein–protein interactions on a genomic scale.

Single chain threading has been widely used for protein tertiary structure prediction, with some success.[28–34] Here, one attempts to align the sequence of the protein of interest to a library of known folds and find the closest matching structure. The goal of threading is to extend sequence-based approaches by recognizing the structures that can be analogous (i.e., the two proteins are not necessarily evolutionary related), but they adopt similar structures; recently, threading has begun to reach this objective.[30]

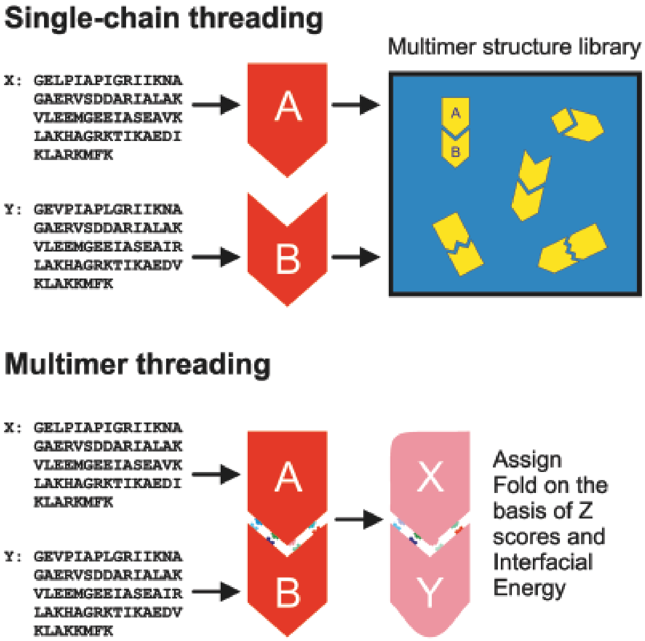

Fig. 1. Principles and strategy of MULTIPROSPECTOR.

To explore our idea that the interactions underlying tertiary and quaternary structure formation are similar, here we propose a novel structure-based approach for protein–protein interaction prediction, MULTIPROSPECTOR, which extends existing threading approaches to multimeric threading; this approach in principle addresses the limitations of existing approaches described above. The qualitative idea of the method is as follows (see Fig. 1). First, we thread the sequences through a representative structure template library that, in addition to monomers, also includes each of the chains in representative protein dimer structures. (The methodology described below is intended to be applicable to an arbitrary number of interacting chains, but in this first application, we focus on dimers for simplicity.) By incorporating knowledge-based, statistical interfacial pair potentials, we then compute the interaction energy between a pair of protein chains for those protein structures involved in dimeric complexes. Whether two proteins in fact form a stable complex is determined by the magnitude of the interfacial potentials and the Z-scores of the complex structures relative to that of the monomers. Although the method depends on the presence of solved complex structures, the query sequences do not need to have solved structures, and the results automatically reveal the binding sites of the interacting protein pairs.

The organization of this article is as follows. In Materials and Methods, we describe the potentials used in multimeric threading and then the strategy for multimeric threading and the test sets to be used to assess this approach. In Results, we present findings on the evaluation of multimeric threading on a benchmark consisting of 69 monomers, 40 homodimers, and 15 heterodimers. The fold library that we thread against consists of 2478 folds; the sequence identity between each of two folds in the threading template library is <35%. The criteria for assigning proteins to a multimer are also described. Then, we predict a number of interacting proteins in yeast and compare the results with known data. The resulting quaternary structures may be found on our Web site (http://bioinformatics.danforthcenter.org/services/proint/). And the estimation of the false negative rate is provided. We also compare the performance of MULTIPROSPECTOR with the assignments made by PSI-BLAST.[35] Finally, in Discussion, we summarize the present work, highlight the significance of this approach, and discuss the limitations and envisioned future improvements.

MATERIALS AND METHODS

DIMER: A Database of Dimer Template Structures

The Protein Data Bank (PDB) stores three-dimensional structures of macromolecules, some of which are cocrystallized proteins.[36] Our dimer database, DIMER, is constructed by selecting the cocrystallized records from the PDB with use of the criteria listed below:

1. The resolution of the two-chain PDB records should be ≤2.5 Å.
2. The threshold for the number of interacting residues is set to be >30 to avoid crystallizing artifacts. Interacting residues are defined as a pair of residues from different chains that have at least one pair of heavy atoms within 4.5 Å of each other.
3. Each chain in the dimer database should have >30 amino acids to be considered as a domain.
4. Dimers in the database should not have >35% identity with each other (i.e., at most, one chain in a complex can have >35% identity to any of the chains in another complex.
5. The dimers should be confirmed in the literature as genuine dimers instead of crystallization artifacts.

This selection results in 340 dimers, including 271 homodimers and 69 heterodimers. To construct a threading template library representing each chain in the dimer databases, as well as other monomer structures, for single chain threading, PDB structures are clustered according to sequence identity criteria (35%). Within each cluster, only one protein is chosen to represent the cluster. When chains from the 340 dimers appear in the cluster, we select one dimer chain to represent the cluster. In this way, a representative set of PDB structures is selected as our single-chain-threading template library, which contains 2478 folds where the sequence identity between each two folds is <35% to avoid the possible problems of overfitting. These 2478 templates are composed of 268 chains from homodimers, 96 chains from heterodimers, and 2114 chains from monomers or possibly higher order multimers. This fold library can be found on our Web site at http://bioinformatics.danforthcenter.org/services/point/.

Interfacial Statistical Potentials Used for Multimeric Threading

The statistical interfacial pair potentials are developed from the dimer database, DIMER, as described above. The interfacial pair potentials, P(i, j), (i = 1, …, 20; j = 1, …, 20), are calculated by examining each interface of the selected dimers with use of the following formula:

$$P(i,j) = -\log\left(\frac{N_{\text{obs}}(i,j)}{N_{\text{exp}}(i,j)}\right) \tag{1a}$$

where $N_{\text{obs}}(i,j)$ is the observed number of interacting pairs of i,j between two chains. $N_{\text{exp}}(i,j)$ is the expected number of interacting pairs of i,j between two chains if there are no preferential interactions among them. The expected number can be calculated from:

$$N_{\text{exp}}(i,j) = X_i \times X_j \times N_{\text{total}} \tag{1b}$$

where X_i is the mole fraction of residue i in total surface residues. N_{total} is the number of total interacting pairs. The definition of interacting pairs is the same as in the dimer database selection (i.e., a pair of residues from different chains that have at least one pair of heavy atoms within 4.5 Å of each other).

By applying Boltzmann's principle to the ratio of the observed frequencies to expected frequencies of pairings between two residue types, one obtains an estimate of the potential of mean force between those two residue types. The details of the construction and evaluation of the current potential and other types of statistical potentials are shown elsewhere (manuscript in preparation), but the potential itself can be found on our Web site at http://bioinformatics.danforthcenter.org/services/proint/.

Multimeric Threading Protocol

The multimeric threading approach of MULTIPROSPECTOR is illustrated in Figure 2 and consists of two phases. Phase I involves single-chain threading, where each sequence is independently threaded and assigned a list of possible candidate structures (note that a permissive Z-score cutoff is used so that sequences that weakly prefer monomers but strongly prefer multimers are not missed). Phase II uses multi-chain threading, where a set of probe sequences, each at least weakly assigned to a monomer template structure that is part of a complex, is then threaded in the presence of each other in the associated quaternary structure. If the interfacial energy and Z-score are sufficiently favorable, then the sequences are assigned this quaternary structure.

In phase I, we use our threading program PROSPECTOR (Protein Structure Predictor Employing Combined Threading to Optimize Results). PROSPECTOR has been described elsewhere in detail; here we briefly summarize the methodology.[37] First, both close sequence profiles (whose pairwise identify is between 35 and 90%) and distant sequence profiles (all pairs of sequences having an E-value < 10) are generated. Then, each of these sequence profiles is used to scan a structural database. The probe-template alignments provided by the sequence profile scoring function are used to identify the partners in the probe sequence for use in the next threading iteration that uses sequence plus secondary structure plus pair interaction profiles. The five top-scoring structures for each scoring scheme are collected and composite results are reported. To measure the significance of the alignments, PROSPECTOR gives a scoring function that is related to the sequence alignment and pairwise interactions. The

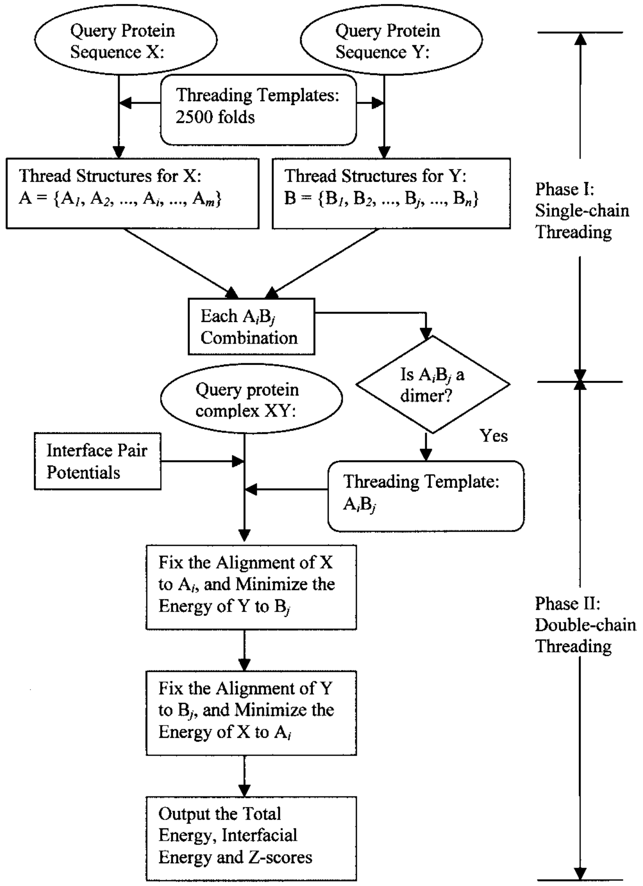
Query Protein Sequence X:
Query Protein Sequence Y:
Threading Templates: 2500 folds
Thread Structures for X: A = {A1, A2, ..., Ai, ..., Am}
Thread Structures for Y: B = {B1, B2, ..., Bj, ..., Bn}
Phase I: Single-chain Threading
Each AiBj Combination
Is AiBj a dimer?
Yes
Query protein complex XY:
Interface Pair Potentials
Threading Template: AiBj
Fix the Alignment of X to Ai, and Minimize the Energy of Y to Bj
Phase II: Double-chain Threading
Fix the Alignment of Y to Bj, and Minimize the Energy of X to Ai
Output the Total Energy, Interfacial Energy and Z-scores

Fig. 2. Illustration of our multimeric threading strategy. The protocol for MULTIPROSPECTOR is comprised of two phases. In phase I, both sequences X and Y are independently threaded by using PROSPECTOR. A set of templates A and B with initial Z-score > 2.0 is identified. Phase II begins with the decision of whether the template structure pair A_iB_j is part of a known complex. Only when A_iB_j forms a complex does multimeric threading continue to rethread on the partners in the complex and incorporate the protein–protein interfacial energies. Double-chain threading is used in this step. It first fixes the alignment of X to the template A and adjusts the alignment of Y to the template B, and then it fixes the alignment of Y to the template B and adjusts the alignment of X to the template A. Finally, the algorithm gives the template A_iB_j that has the highest Z-score as a possible solution. At the same time, the algorithm provides the total energy of the complex as well as the interfacial energy.

Z-score of the score for each probe-template alignment is used to decide if a correct fold is found:

$$Z_K = \frac{E_K - \langle E \rangle}{D}, \tag{1c}$$

where E_k is the energy of the Kth fold,

$$\langle E \rangle = \frac{1}{M}\sum_{i=1}^{M} Ei, \tag{1d}$$

is the average energy, and

$$D = \left[\frac{1}{M}\sum_{i=1}^{M} Ei - \langle E \rangle)^2\right] \tag{1e}$$

is the standard deviation of energies; Ei is the energy of the i-th sequence of M alternative folds ($i = 1, \ldots, M$). The Z-score gives the average number of standard deviations between the Kth and the random fold energy.[37] The confident threshold for fold assignments is empirically found to be a Z-score above 5.0 (good Z-scores are positive).

During phase I, PROSPECTOR independently threads both sequences X and Y. Then, those probe sequences aligned to templates with a Z-score >2.0 are identified. Suppose A is a set of templates selected by sequence X, and B is a set of templates selected by sequence Y. Thus, a pool of possible structures A and B is generated by PROSPECTOR, where A = {A_1, A_2, ..., Ai, ..., Am} and B = {B_1, B_2, ..., Bj, ..., Bn}. The Z-score of each alignment, as well as the initial alignment, is provided by PROSPECTOR after phase I.

Phase II begins with the decision of whether the combination of template structures A_i and B_j forms a protein–protein complex, that is, whether A_iB_j is in the dimer database. Only when A_iB_j forms a complex does multimeric threading occur. The next step is to use a double-chain-threading algorithm to thread the composite sequences XY together on the composite template A_iB_j. First, the initial alignment of X to A_i, generated by PROSPECTOR in phase I, is fixed. Then, the alignment between Y and B_j by incorporating the statistical interfacial pair potentials to optimize the energy between Y and B_j is adjusted. Next, the adjusted alignment of Y to B_j is held fixed, and the alignment between X and A_i is modified as was the alignment between Y and B_j in the previous round. In principle, this should be repeated a number of times and be independent of the order, but here we consider this very simple case. Next, the algorithm gives template A_IB_J with the highest Z-score as a possible solution. At the same time, the algorithm provides the total energy of the complex as well as the interfacial energy.

Whether A_IB_J is the predicted quaternary structure depends on three possible outcomes. First, if either query sequence X or Y (or both) hits another monomer template C with a much more significant score (greater Z-score) than any template structure that forms a dimer, we assign the query sequence to be a monomer with the structure of template $_C$. Second, if both template structures A_I and B_J were initially predicted with sufficient confidence (Z-score > 5.0) in phase I, then the binding prediction is based on the interfacial energy evaluated with knowledge-based potentials. Only if the energy is lower than a certain threshold E_0, based on known complexes (see below), are these two chains predicted to interact. Otherwise, they are assigned to the monomeric structures A_I and B_J. Third, if neither or only one of the Z-scores in the initial alignments is >5.0, we require both of the Z-scores to be above 5.0 after multimeric threading. At the same time, the interfacial energy should be lower than E_0 before the prediction is made.

The logic of this set of rules is that if the folds of the individual proteins are confidently assigned (Z initial ≥ 5.0), then we only need to check if these folds form a complex in our database. If no monomer fold is confidently assigned, we will check if mediumly confident folds (2.0 ≤ Z initial < 5.0) form a complex, in which case we would

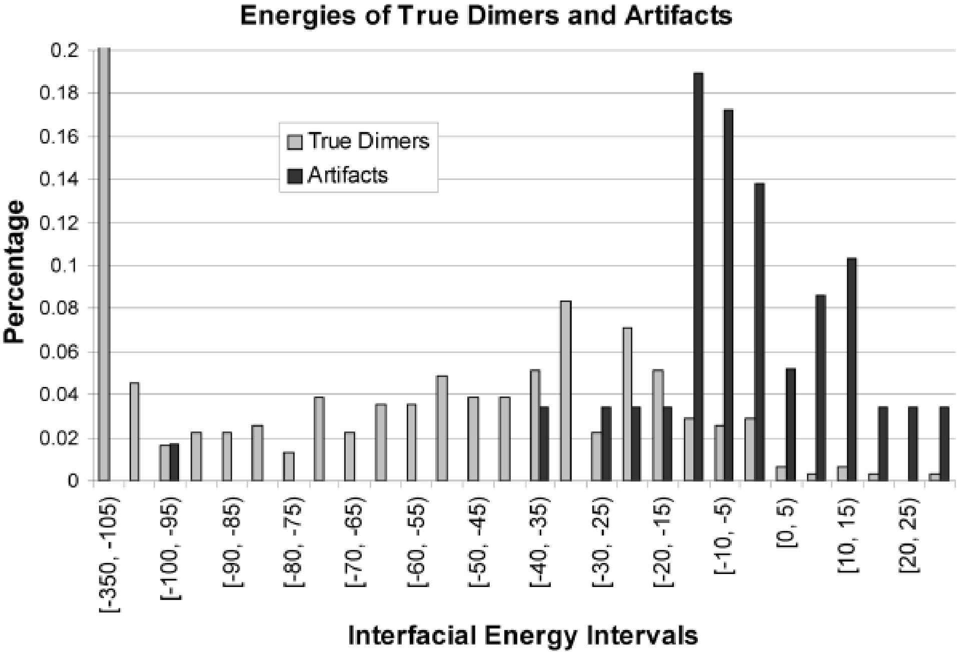

Fig. 3. The determination of the threshold of interfacial energy, E_0. Statistical interfacial potentials are applied on a true dimer test set and a set of monomers that do not form true dimers but cocrystallize to determine a threshold of the interfacial energy that can distinguish between true dimers and cocrystallized artifacts that results in the minimum number of false positives and false negatives. The results are shown below. The number of cases in each energy interval has been normalized by the total number of proteins in the test sets. A threshold of -15.0 is consequently set from this histogram.

require that after multimeric threading the Z-score will improve to above 5.0. The threshold of energy E_0 is set to be -15.0 to make sure that the interfacial interactions are strongly favored (good interfacial energies are negative). The process of construction of the potentials and determining the threshold of interfacial energy E_0 will be discussed in detail in another manuscript that is currently in preparation. To briefly summarize, the threshold was determined as follows. A test set of 310 true dimers and a test set of 58 monomers, which are different from the test cases shown in this article, are selected. The potentials are applied on these sets and a threshold (-15) that results in the maximum number of true positives and true negatives was determined (see Fig. 3).

Test Cases

A set of homodimers and monomers was chosen from the data in an article by Ponstingl et al.[38] The data set is comprised of 96 monomers and 76 homodimers. The protein chains within each subset exhibit <25% sequence identities and are structurally dissimilar. A set of heterodimers is selected from the data in an article by Norel et al.[25] and is comprised of structures identified by 26 PDB codes. We noticed that some of the dimers in these datasets consist of multiple chains. The PDB records that do not contain two chains are subsequently removed. The PDB records, which have identical counterparts in our dimer database, are also removed, resulting in 69 monomers, 40 homodimers, and 15 heterodimers. This list of test proteins is found in Table II. Each of these protein complexes was tested by the multimeric threading algorithm; see RESULTS. Each partner of the dimers was independently threaded by using PROSPECTOR through the fold library having 2478 structures in phase I. Then, both partners were threaded through our dimer database in phase II.

Prediction of Protein–protein Interactions and Quaternary Structure in Yeast

We downloaded 2457 unique physical interactions that involve 1872 proteins from the MIPS, http://mips.gsf.de/proj/yeast/tables/interaction/physical_interact.html.[39] We then threaded against the entire fold library and applied the entire protocol of MULTIPROSPECTOR to see how many of these “known interactions” can be predicted without any a priori information. Of course, because the dimer database is incomplete, it is impossible to predict protein interactions for pairs involved in complexes whose structure is not yet solved.

We also tried to see how many interactions could be predicted among all 6146 proteins encoded by yeast genome. The Database of Interacting Proteins (DIP) is downloaded from http://dip.doe-mbi.ucla.edu/ to validate our predictions. The comparison is as follows: if both proteins of one prediction have homologs to a pair of proteins in the DIP, this prediction is more likely to be a true interaction than those that do not have counterparts in the DIP. The homology is defined as an E-value <0.01 by PSI-BLAST. Because >70% of the interactions in the DIP were determined by yeast two-hybrid screening, which might introduce potentially high false positives, the valid-

TABLE I. Representative Predictions Made by MULTIPROSPECTOR

	Query/CATH fold[a]	Category[b]	Template/CATH fold[c]	Seqid[d] (%)	Z_x^e	Z_y^f	E[g]	$Z_{x'}^h$	$Z_{y'}^l$	Prediction[j]
1	1amk/3.20.20.90	Hm	1cil/3.20.20.90	68.0	10.7	11.0	−51.5	11.6	11.8	D
2	1bjf/1.10.238.40	Hm	1alv/1.10.238.40	14.5	3.7	3.2	−56.0	5.1	5.3	D
3	1flp/1.10.490.10	Mm	1hbi/1.10.490.10	21.4	5.8	5.6	−12.5	5.9	5.7	M
4	8paz/2.60.40.420	Mm	1ac6/2.60.40.10	1.9	2.1	2.0	−18.6	2.5	3.3	M
5	1tgs/E:5.1.40.1	Ht	1ppf/E:5.1.40.1	33.2	7.1	4.3	−73.4	7.9	5.6	D
	I:2.20.34.12		I:2.20.34.12	25.4						
6	1tec/E:3.40.50.200	Ht	1a10/E:3.40.50.200	44.1	10.7	5.4	−62.0	11.7	5.9	D
	I:3.30.10.10		I:3.30.10.10	35.9						
7	1bkz/2.60.120.60	Mm	1a78/2.60.120.60	30.6	6.7	6.9	−10.3	7.0	7.3	M
8	1slt/2.60.120.60	Hm	1a78/2.60.120.60	48.1	6.9	7.1	−39.8	7.9	7.7	D

[a]PDB codes of the query proteins and their corresponding CATH assignments.

[b]The true biological oligomerization states of the query proteins. Hm, homodimer; Ht, heterodimer; Mm, monomer.

[c]PDB codes of the template proteins that are found to be structural homologues to the query proteins by our threading algorithm, PROSPECTOR, and their corresponding CATH assignments.

[d]Sequence identities between the query proteins and their corresponding template proteins found by PROSPECTOR.

[e,f]Z_x and Z_y are the Z-scores of two query chains calculated after phase I threading.

[g]The interfacial energies between the two query chains calculated by MULTIPROSPECTOR.

[h,i]$Z_{x'}$ and $Z_{y'}$ are the Z-scores of two query chains calculated after phase II.

[e]The oligomerization states assigned by MULTIPROSPECTOR. D, dimer; M, monomer.

ity of the interactions in the DIP itself remain open to questioning.[40,41] However, this is the best way we can think of to validate our results.

Comparison to Simple Approach Using PSI-BLAST

We compared our predictions with a simple approach using PSI-BLAST after assigning structures to the yeast-interacting proteins. The strategy of this simple approach is that if two query proteins X and Y have significant PSI-BLAST hits to A and B that are two chains in a dimer in our database, then we assign X and Y as interacting partners and take the alignment from that of the isolated chains (PSI-BLAST has no structural information about the interacting region). Obviously, this approach depends on the permissiveness of the E-value thresholds. E-values of 1 and 0.01 were examined in the cases discussed below. The PSI-BLAST program was downloaded from NCBI, http://www.ncbi.nlm.nih.gov/BLAST/ and run by following the instructions in the manual.

RESULTS

Examples of Multimeric Threading Prediction

The testing cases for the multimeric threading algorithm should be the proteins or protein complexes with solved structures so that we are able to compare the structures that our prediction assigned with the native structures. To decide if two proteins have the same fold, we initially used the structural classification made by CATH.[42] Several examples are presented here to show how our multimeric threading protocol works under different situations to predict whether two proteins interact or not. To help the reader understand how this multimeric threading algorithm functions, some cases are discussed in detail with the results summarized in Table I and with the full set presented in Table II.

First, we show how real dimers are recognized by multimeric threading by using two homodimers 1amk (triosephosphate isomerase) and 1bjf (neurocalcin delta) as examples. For example, in phase I single chain threading, 1amk has been predicted by PROSPECTOR (Protein Structure Predictor Employing Combined Threading to Optimize Results) to have the same fold as 1ci1 (triosephosphate isomerase) with high confidence (i.e., a Z-score > 5.0. In this case, the only criterion to determine if 1amk is a homodimer is if the interfacial energy is below the threshold energy E_0 (more negative energies are more favorable). Here the interfacial energy is −51.5, well below the energy threshold of −15.0; thus, 1amk is assigned to be a dimer.

In contrast to 1amk, 1bjf shares the same fold as both chains of 1alv (calcium-bound domain VI of procine calpain) in the dimer library, but the single-chain Z-scores assigned by PROSPECTOR are 3.7 and 3.2, respectively, which are below the confident threshold. After applying our second phase of multimeric threading, the Z-scores for both chains improved to be above the confident threshold. The interfacial energy is well below the threshold (Table I). This example shows that by incorporating the interfacial potentials, multimeric threading could improve the sensitivity of the single-chain-threading algorithm.

Next, we show two examples, 1flp (monomeric hemoglobin I) and 8paz (oxidized native pseudoazurin), that show that the real monomers can be correctly assigned by multimeric threading. Both proteins have been confirmed as monomers.[43,44] 1flp has a significant hit to 1hbi (homodimeric hemoglobin I), a homodimer in our threading template database with Z-scores of 5.8 and 5.6 for each respective chain. After running multimeric threading, the interfacial energy is −12.5, which is above the energy threshold of −15. Thus, 1flp is correctly assigned to be a monomer. This example shows that although 1flp and 1hbi have the same fold by CATH, our algorithm can still successfully differentiate them by examining the interfacial energy.

TABLE II. Test Data Set of Monomers, Homodimers, and Heterodimers[†]

Monomers			
16PK kinase	1A0K cytoskeleton	1A8O capsid	1AAY complex (zinc finger/DNA)
1AFK hydrolase	1AKZ glycosidase	1AM6 hydrolase	1AMJ lyase(carbon-oxygen)
1AYI bacteriocin	1AYL kinase (transphosphorylating)	1BC2 hydrolase	1BE0 dehalogenase
1BEA serine protease inhibitor	1BGC cytokine	1BKZ lectin	1BMB hormone/growth factor
1BRY ribosome-inactivating protein	1BU1 transferase	1CTJ electron transport	1DFF hydrolase
1DJX lipid degradation	1DMR oxidoreductase	1EMA fluorescent protein	1ESF enterotoxin
1ESO oxidoreductase	1FDR flavoprotein	1FEH oxidoreductase	1FLP oxygen transport
1GCI serine protease	1IAE zinc endopeptidase	1IPS antibiotic biosynthesis	1KFS complex (hydrolase/DNA)
1KPT toxin	1MB1 transcription regulation	1MDT toxin	1MH1 GTP-binding
1MPG hydrolase	1NP4 transport protein	1NUC nuclease	1OPS antifreeze protein
1PDA lyase (porphyrin)	1PPO hydrolase (thiol protease)	1RGP G-protein	1RHS transferase
1TON hydrolase (serine proteinase)	1UCH cysteine protease	1VJW oxidoreductase	1XGS aminopeptidase
1YGE dioxygenase	1ZIN phosphotransferase	232L hydrolase	2ABX postsynaptic neurotoxin
2ACY acylphosphatase	2ATJ oxidoreductase	2BLS cephalosporinase	2CY3 electron transport (heme protein)
2END endonuclease	2FGF growth factor	2GPR phosphotransferase	2HEX blood clotting
2IHL hydrolase (o-glycosyl)	2MBR oxidoreductase	2MHR oxygen binding	2RN2 hydrolase (endoribonuc lease)
3CMS hydrolase (acid proteinase)	3DFR oxido-reductase	3SIL glycosidase	5CP4 oxidoreductase
8PAZ electron transfer			
Homodimers			
1A3C transcription regulation	1AJS aminotransferase	1ALO oxidoreductase	1AMK gluconeogenesis
1AOM oxidoreductase	1BIF bifunctional enzyme	1BSR hydrolase (phosphoric diester, RNA)	1CP2 oxidoreductase
1CSH lyase (oxo-acid)	1DAA transferase (aminotransferase)	1FIP DNA-binding protein	1FRO lactoylglutathione lyase
1GVP DNA-binding protein	11CW cytokine	11MB hydrolase	1ISA oxidoreductase (superoxide acceptor)
1ISO oxidoreductase	1KPF protein kinase inhibitor	1LYN fertilization protein	1MJL transcription regulation
1NOX flavoenzyme	10AC oxidoreductase	10PY isomerase	10TP phosphorylase
1PGT transferase	1RFB glycoprotein	1SLT lectin	1SMN endonuclease
1TOX toxin	1TRK transferase (ketone residues)	1TYS transferase (methyltransferase)	1UTG steroid binding
1WGJ hydrolase	1XSO oxidoreductase (superoxide acceptor)	2CCY electron transport (heme protein)	2RSP hydrolase (aspartyl proteinase)
3GRS oxidoreductase (flavoenzyme)	3SDH oxygen transport	5TMP transferase	9WGA lectin (agglutinin)
Heterodimers			
1BZX trypsin/trypsin inhibitor complex	1CGI serine proteinase/ inhibitor complex	1CHO serine proteinase/ inhibitor complex	1CSE serine proteinase/inhibitor complex
1HWM hydrolase (ebulin)	1OUT hemoglobin	1PBX hemoglobin	1RSC lyase (carbon-carbon)
1SBN subtilisin/eglin complex	1TEC thermitase/eglin complex	1TGS trypsinogen/trypsin inhibitor complex	2PTC proteinase/inhibitor complex
2SNI subtilisin/chymotrypsin inhibitor complex	2TGP trypsinogen/ trypsin inhibitor complex	4TPI proteinase/inhibitor complex	

[†]A benchmark set composed of 69 monomers, 40 homodimers, and 15 heterodimers are listed. Each protein-protein complex is represented by their PDB code followed by their functional annotations.

8paz does not have a significant hit in our dimer database but has a medium confident hit (i.e., a Z-score > 2.0), 1ac6 (T-cell receptor alpha). After running multimeric threading, the Z-score of neither chain was increased to >5.0. Hence, 8paz is again correctly classified as a monomer. Although it is possible that an unknown protein may still interact without adopting any known templates, 8paz has a significant hit to 1ag6 (plastocyanin), a monomer in our threading template database; thus, it is likely that 8paz exists as a monomer.

Considering that the binding of heterodimers (where each of the two chains has <35% sequence identity) may have different binding mechanisms than homodimers, we also tested the performance of our multimeric threading on heterodimers. 1tgs (proteinase/inhibitor complex) and 1tec (proteinase/inhibitor complex) are selected as examples.[45,46] 1tgs has one chain with a Z-score <5.0, whereas both chains of 1tec have Z-scores >5.0. After multimeric threading, the Z-score of the medium confident chain in 1tgs increased to 5.6. Both of the complexes have favorable energies which are lower than −15.0. Thus, both are correctly assigned as dimers. These examples show that MULTIPROSPECTOR is also able to predict interactions in heterodimers.

Another possible concern with this or in fact any multimeric threading algorithm is that it may not be able to differentiate between proteins that have the same fold but adopt different oligomerization states. To test whether this is true, we selected 1bkz (galectin-7) and 1slt (S-lectin) as examples (Fig. 2); both have significant Z-scores as shown in Table I when threaded to 1a78 (galectin-1) in our dimer database. However, 1bkz has been shown to be a biological monomer, whereas 1slt is a biological dimer.[38] After running multimeric threading, the two proteins have significantly different interfacial energies, with that of 1bkz higher than −15.0, namely −10.3. Thus, 1bkz is classified as a monomer. In contrast, 1slt has an interfacial energy of −39.8 and is classified as a dimer. From this calculation, we can see that by using interfacial energies, different oligomerization states can successfully be distinguished.

Monomer Test Set (69 Cases)

Sixty-nine monomers have been selected for extensive tests by using the current protocol; see Table II. After phase I threading, 44 of them have hits in the dimer library with Z-scores >2.0, with the remainder assigned to monomers. After phase II multimeric threading, for 40 monomers, either the Z-scores of the complex structure did not improve to >5.0 or the interfacial energies do not favor a dimer. Thus, in only four cases has the protein been incorrectly predicted as forming a homodimer.

The four apparently falsely assigned cases are 1ema (green fluorescent protein from *A. victoria*), 1ppo (protease omega), 1ton (tonin), and 2bls (AMPC beta-lacamase). 1ema, a green fluorescent protein, has a hit to 1gfl (green fluorescent protein from *A. victoria*) in our dimer database. According to Tsien's recent article, these two proteins might both exist as dimers in vivo.[47] For 1ppo, 1ton, and 2bls, our method fails to predict the two proteins as monomers.

A comparison with the simple approach by PSI-BLAST shows that in addition to the four false predictions by our method, PSI-BLAST falsely assigns another 10 monomers as dimers, namely, 1be0, 1bkz, 1bu1, 1eso, 1fdr, 1flp, 1nuc, 2abx, 3dfr, and 5cp4. Cases 1be0, 1nuc, 3dfr, and 5cp4 have been assigned as monomers by our method by confident hits to monomer templates, whereas 1bkz, 1eso, 1fdr, and 1flp have been assigned monomers because of unfavorable interfacial energies. 1bu1 does not have a hit in our dimer templates, and the Z-scores of 2abx in phase II did not surpass 5.0.

Homodimer Test Sets (40 Cases)

Forty homodimers have been checked with the current procedure; see Table II. Table III lists the results of each case. These 40 test cases have various sequence identities to the template dimers. In 14 cases, each of the query sequences hits a template with <35% identity. The high-sequence identity cases are less difficult to predict, but because this is the first presentation of our protocol, we did not abandon these cases. More importantly, unknown proteins with high-sequence identity to the known interacting proteins do not guarantee that those unknown proteins also interact, because it is not uncommon that mutations on specific sites disrupted protein interactions.[48] After phase I threading, 39 of these 40 homodimers have at least one hit to the complexes in the dimer library with a Z-score > 2.0. For 38 of these 39 cases, after phase II multimeric threading, the Z-scores of both chains increased to >5.0, and the interfacial energies are lower than −15.0, which means that the dimer is predicted. The one protein whose improved Z-score does not surpass 5.0 is 9wga (wheat germ agglutinin). The protein that has no hit in our dimer database is 1tox (diphtheria toxin). By examining the PDB structures of these two proteins, we found that both 9wga and 1tox have <30 pairs of interactions on the interface. The templates they could have hit in the dimer database had been excluded because they have too few interfacial contacts. Careful examinations on each query and template pair reveal two more false predictions. 1alo (oxidoreductase) hits to the template of 1fo4 (dehydrogenase) in phase I. After phase II multimeric threading, the Z-scores of both chains increased to >5.0, and the interfacial energies are < −15.0. However, 1alo and 1fo4 do not share the same fold assigned by CATH; therefore, even though 1alo is correctly predicted as dimer, we still cannot assign the structure of 1fo4. A similar case is 1imb (hydrolase) and its template, 1bfl (hydrolase). These two false cases show that if single-chain threading gives a wrong prediction, the dimer prediction can be misleading.

The simple approach by PSI-BLAST not only does not predict the four cases missed by our method but also introduces four additional false predictions: 1icw, 1nox, 2ccy, and 2rsp. PSI-BLAST is unable to recognize these cases because of their low-sequence identities to the corresponding dimer templates.

TABLE III. Test Predictions on Homodimers

	Query[a]	Template[b]	Seqid[c]	$Z_{x(y)}$[d]	$Z_{x'(y')}$[e]	E[f]
1	1A3C	1A4X	95.9	7.2	7.8	−37.2
2	1AJS	1AHE	38.1	9.2	10.8	−153.9
3	1ALO	1FO4	19.1	7.9	8.1	−31.7
4	1AMK	1CI1	68.0	10.8	11.6	−59.1
5	1AOM	1AOF	82.5	13.2	13.4	−36.2
6	1BIF	1FBT	36.8	6.3	6.6	−34.2
7	1BSR	1A2W	81.5	9.4	11.3	−75.3
8	1CP2	1DE0	65.1	11.7	12.0	−17.0
9	1CSH	5CSC	92.4	8.2	10.5	−176.1
10	1DAA	1A0G	98.6	9.1	11.4	−145.0
11	1FIP	1ETK	84.8	4.2	7.0	−124.6
12	1FRO	1F9Z	28.7	8.0	11.0	−141.5
13	1GVP	1YHA	98.9	4.5	6.2	−64.8
14	1ICW	1CM9	16.4	2.8	5.1	−61.1
15	1IMB	1BFL	2.2	3.2	8.7	−80.7
16	1ISA	1ABM	41.7	17.1	18.8	−60.5
17	1ISO	1CM7	24.9	8.6	10.5	−112.6
18	1KPF	1AV5	98.2	7.4	9.4	−112.3
19	1LYN	3LYN	66.4	8.2	8.9	−33.7
20	1MJL	1CMB	99.0	5.4	8.6	−122.9
21	1NOX	1BKJ	23.6	4.9	10.5	−149.9
22	1OAC	1KSI	26.1	9.9	11.3	−123.4
23	1OPY	1E3R	98.0	6.5	8.3	−95.0
24	1OTP	1BRW	42.4	10.5	11.2	−48.5
25	1PGT	1GSR	82.4	4.3	7.3	−75.4
26	1RFB	1D9C	98.3	7.0	10.4	−187.4
27	1SLT	1A78	48.1	7.0	7.8	−39.8
28	1SMN	1QAE	99.6	8.0	9.2	−76.6
29	1TOX	No hit	/	/	/	/
30	1TRK	1AY0	99.6	22.8	25.2	−98.5
31	1TYS	1BKP	35.5	7.4	8.0	−64.3
32	1UTG	2UTG	55.7	4.6	6.7	−83.3
33	1WGJ	1IPW	16.5	7.4	8.2	−57.6
34	1XSO	1AZV	66.0	12.6	13.4	−28.6
35	2CCY	1BBH	25.9	6.5	7.8	−67.0
36	2RSP	1BDR	21.2	3.5	6.4	−180.9
37	3GRS	1AOG	26.4	11.1	14.6	−146.2
38	3SDH	1HBI	20.1	5.9	6.2	−31.7
39	5TMP	1TMK	27.8	6.5	7.9	−80.6
40	9WGA	1KBA	7.8	2.0	4.5	−55.0

[a] PDB codes of the query proteins.
[b] PDB codes of the template proteins that are found to be structural homologs to the query proteins by our threading algorithm, PROSPECTOR.
[c] Sequence identities between the query proteins and their corresponding template proteins found by PROSPECTOR.
[d] $Z_{x(y)}$ is the Z-score of the query chain calculated after phase I.
[e] $Z_{x'(y')}$ is the Z-score of the query chain calculated after phase II.
[f] The interfacial energies between the two query chains calculated by MULTIPROSPECTOR.

Heterodimer Test Sets (15 Cases)

Fifteen cases have been selected to test our prediction protocol. Similar to the homodimer test set, we considered test cases with various degrees of sequence identity. Table IV lists the results of each case. In eight cases, the original single-chain Z-score is >5.0, and binding energy is < −15.0. In the other cases, the complex Z-score increased to >5.0, and the binding energy is favorable for dimer formation. Thus, all 15 are correctly assigned.

TABLE IV. Test Predictions on Heterodimers

	Query[a]	Template[b]	Seqid[c]	$Z_{x(y)}$[d]	$Z_{x'(y')}$[e]	E[f]
1	1choE	1ppfE	30.6	7.4	8.6	−73.8
	1choI	1ppfI	100.0	4.8	5.6	
2	1tecE	1a10E	44.1	10.2	10.9	−62.0
	1tecI	1a10I	35.9	5.4	5.9	
3	1tgsZ	1ppfE	33.2	7.0	7.9	−73.4
	1tgsI	1ppfI	25.4	4.2	5.7	
4	2ptcE	1brcE	73.1	9.8	11.5	−70.0
	2ptcI	1brcI	44.1	4.3	5.3	
5	2sniE	1a10E	69.5	9.7	9.9	−54.8
	2sniI	1a10I	96.9	5.1	6.1	
6	2tgpZ	1brcE	73.1	9.8	11.5	−70.0
	2tgpI	1brcI	44.1	4.3	5.3	
7	1cseE	1a10E	99.3	7.8	8.7	−53.4
	1cseI	1a10I	35.9	5.4	5.9	
8	1cgiE	1ppfE	29.9	7.1	8.4	−72.9
	1cgiI	1ppfI	29.8	4.2	5.3	
9	1bzxE	1brcE	65.9	10.7	11.7	−73.5
	1bzxI	1brcI	44.1	4.4	5.3	
10	1rscA	1ausL	77.7	9.1	9.3	−49.5
	1rscM	1ausS	40.3	8.4	9.2	
11	1outA	1fdhA	57	9.2	9.6	−16.1
	1outB	1fdhG	51.4	9.3	9.6	
12	1sbnE	1a10E	69.5	11.9	12.8	−39.6
	1sbnI	1a10I	35.9	5.3	5.5	
13	1pbxA	1fdhA	49.3	9.3	9.7	−18.6
	1pbxB	1fdhG	45.9	9.4	9.7	
14	1hwmA	1abrA	39.1	8.7	9.8	−109.9
	1hwmB	1abrB	44	12.8	14.6	
15	4tpiZ	1brcE	73.1	10.5	11.3	−69.6
	4tpiI	1brcI	45.8	4.0	5.1	

[a] PDB codes of the query proteins.
[b] PDB codes of the template proteins that are found to be structural homologs to the query proteins by our threading algorithm, PROSPECTOR.
[c] Sequence identities between the query proteins and their corresponding template proteins found by PROSPECTOR.
[d] $Z_{x(y)}$ is the Z-score of the query chain calculated after phase I.
[e] $Z_{x'(y')}$ is the Z-score of the query chain calculated after phase II.
[f] The interfacial energies between the two query chains calculated by MULTIPROSPECTOR.

The simple approach by PSI-BLAST also correctly assigned these 15 cases. More heterodimers are probably needed to differentiate the multimeric threading and the PSI-BLAST approaches. However, it is hard to find the heterodimer test cases with low-sequence identity to the template but with the same fold as the template.

The true and false predictions of the above three test sets have been plotted in Figure 4. The total success rate is >90%, and both the false-positive and the false-negative rates are estimated <10%.

Application to Yeast-Interacting Proteins

We have applied our protocol to the yeast-interacting proteins downloaded from MIPS. There are 2457 unique interactions involving 1872 yeast proteins. All of these interactions have been identified by experimental techniques, such as yeast two-hybrid screening. MULTIPROSPECTOR has been applied to all 1872 proteins. When two proteins are independently threaded to the two partners of

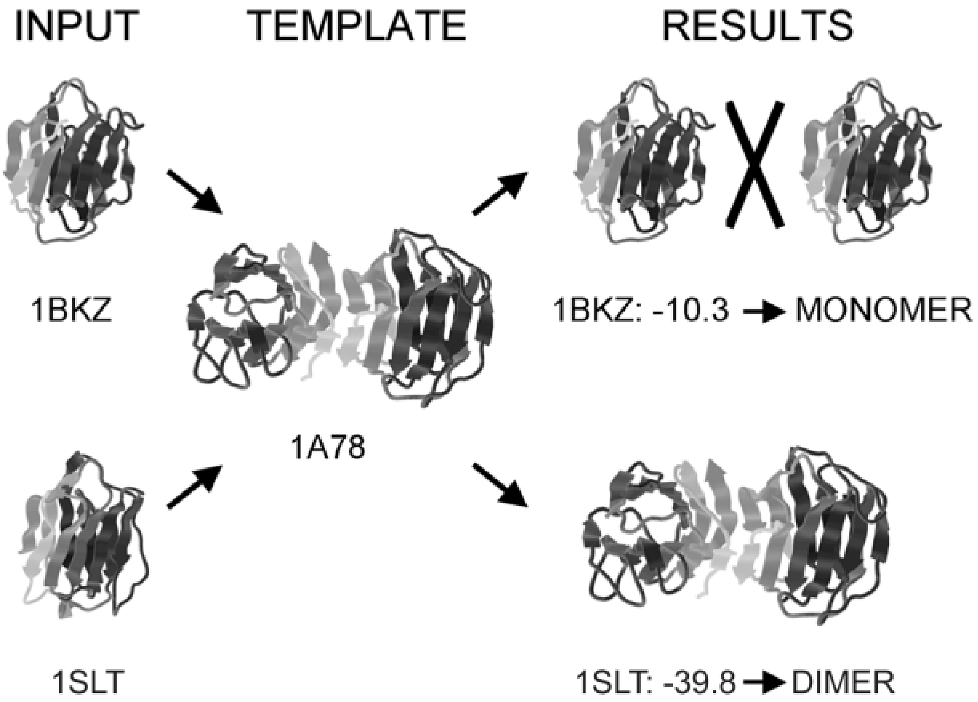

Fig. 4. The illustration of 1bkz and 1slt.

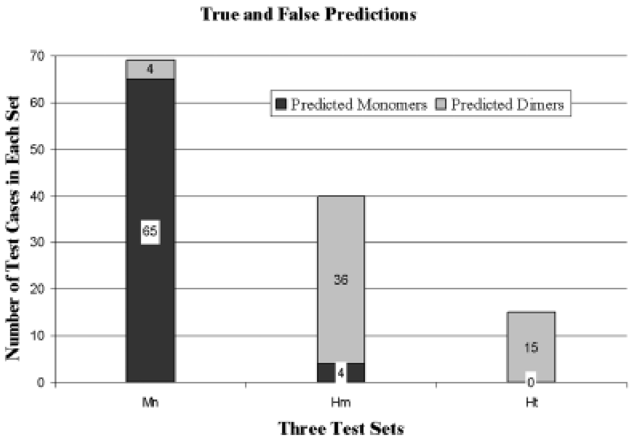

Fig. 5. True and false predictions of the three test sets. Three test sets, 69 monomers, 40 homodimers, and 15 heterodimers, have been selected to test our method. Our method correctly recognized and assigned 36 homodimers, 15 heterodimers, and 65 monomers. Mn stands for monomers, Hm stands for homodimers, and Ht stands for heterodimers.

a dimer with medium confidence (Z-score > 2.0), a pair of potential interactions is identified. After phase I threading, 212 possible pair of interactions of the 2457 known interactions have been identified. After phase II multimeric threading, 144 interactions satisfy our criteria (Z-score > 5.0 and interfacial energy < −15) to be dimers. The results are shown in Figure 5 and in Table V. The predicted interactions and their corresponding structures may be found on our Web site, http://bioinformatics.danforthcenter.org/services/proint/. It is worth emphasizing that MULTIPROSPECTOR did not use the information about which proteins interact as input, so the 144 interactions listed in the Web site are true predictions.

In another test, we compared our prediction results with the simple approach using PSI-BLAST described in Materials and Methods. The results are shown in Figure 6 and Table VI as well. The simple approach using PSI-BLAST assigned structures to 56 protein complexes, 25 of which overlap with the predictions made by our method. As we noticed in this prediction, the E-value for PSI-BLAST was set to be 1.0, which is quite permissive. If we decrease the E-value to be 0.01, which is a more conservative (and reasonable) value, then the number of predictions made by the simple approach decreases, resulting in 26 assigned and 18 overlapped with our multimeric threading predictions.

Finally, we applied our protocol to predict the interactions among all 6146 proteins encoded by the yeast genome. In total, 2865 interactions have been assigned multimeric structures by our protocol. Within these, 2721 interactions have not been reported in the MIPS. The predicted interactions and their corresponding structures may be found on our Web site, http://bioinformatics.danforthcenter.org/services/proint/. To validate our predictions, we attempted to compare the results with the Database of Interacting Proteins (DIP), which is a database that documents experimentally determined protein–protein interactions.[49] The current version of the DIP contains approximately 11,000 unique interactions among 5900 proteins from >80 organisms including *S. cerevisiae*.[49] Among these 2865 interactions, 1138 have counterparts in DIP. We also made the protein–protein interaction predictions with the simple approach using PSI-

TABLE V. MULTIPROSPECTOR Predictions of Yeast Interacting Proteins[†]

Predicted pairs		Predicted pairs		Predicted pairs	
YAL001C	YDR362C	YGL178W	YLR452C	YLR319C	YLL021W
YBR133C	YDL154W	YGL207W	YPR135W	YLR347C	YGL092W
YBR133C	YPL016W	YGL238W	YNL236W	YLR347C	YIL115C
YBR202W	YEL032W	YGR014W	YIL144W	YLR347C	YKL068W
YCR005C	YCR005C	YGR061C	YLR386W	YLR347C	YLR335W
YCR077C	YNL088W	YGR119C	YMR308C	YLR347C	YMR047C
YCR092C	YNL082W	YHR102W	YOR353C	YLR438C-A	YER146W
YDL077C	YDR080W	YHR158C	YGR238C	YLR442C	YDR227W
YDL126C	YDL190C	YHR172W	YNL126W	YML065W	YKR101W
YDL132W	YGL249W	YIL026C	YFL008W	YMR032W	YIL159W
YDL132W	YJR090C	YIL026C	YJL074C	YMR080C	YHR077C
YDL154W	YIL144W	YIL074C	YER081W	YMR080C	YJR132W
YDL215C	YPR048W	YIL074C	YIL074C	YMR129W	YML103C
YDR074W	YML100W	YIL109C	YDL195W	YMR231W	YLR396C
YDR074W	YMR261C	YIL115C	YDR395W	YMR231W	YPL045W
YDR085C	YPL242C	YIR006C	YGL094C	YMR261C	YML100W
YDR097C	YNL082W	YJL026W	YGR180C	YNL102W	YGL207W
YDR108W	YDR407C	YJL041W	YDR395W	YNL118C	YMR080C
YDR108W	YMR218C	YJL041W	YJL061W	YNL201C	YPR115W
YDR118W	YKL022C	YJL057C	YMR129W	YNL216W	YDR227W
YDR118W	YOR249C	YJL187C	YBR133C	YNL216W	YDR464W
YDR264C	YDR103W	YJR089W	YGR140W	YNL216W	YLR442C
YDR264C	YPL242C	YJR132W	YIL115C	YNL243W	YNL243W
YDR301W	YLR115W	YJR159W	YDL246C	YNL287W	YBR281C
YDR301W	YLR277C	YKL067W	YKL067W	YOL004W	YDR207C
YDR335W	YGR009C	YKL101W	YBR133C	YOL004W	YHR178W
YDR335W	YPR008W	YLR006C	YCR073C	YOL004W	YMR019W
YDR356W	YHR172W	YLR014C	YLR014C	YOL004W	YMR053C
YDR356W	YNL126W	YLR071C	YNL236W	YOL051W	YNL236W
YDR407C	YMR218C	YLR115W	YLR277C	YOL090W	YCR092C
YDR484W	YJL029C	YLR127C	YDR118W	YOL090W	YDR097C
YDR490C	YLR466W	YLR127C	YKL022C	YOL090W	YNL082W
YEL046C	YEL046C	YLR127C	YOR249C	YOR033C	YOL090W
YEL061C	YEL061C	YLR148W	YLR396C	YOR128C	YOR128C
YER099C	YOL061W	YLR148W	YMR231W	YOR160W	YIL115C
YFL008W	YFR031C	YLR148W	YPL045W	YOR202W	YOR202W
YFL008W	YIL144W	YLR166C	YGL233W	YOR249C	YKL022C
YFL008W	YJL074C	YLR233C	YGL201C	YOR290C	YJL176C
YFR002W	YJL039C	YLR233C	YJL019W	YOR371C	YGR218W
YFR002W	YML103C	YLR233C	YKL020C	YPL045W	YLR396C
YFR031C	YIL144W	YLR245C	YLR245C	YPL075W	YNL216W
YFR037C	YIL126W	YLR274W	YBR202W	YPL147W	YKL188C
YFR037C	YOR290C	YLR274W	YDL132W	YPL153C	YDR217C
YGL016W	YLR335W	YLR274W	YEL032W	YPL174C	YIL144W
YGL019W	YOR039W	YLR274W	YPL160W	YPL174C	YOR361C
YGL094C	YKL025C	YLR275W	YPR182W	YPR032W	YGR009C
YGL145W	YPR105C	YLR310C	YLL016W	YPR135W	YNL102W
YGL173C	YCR077C	YLR310C	YLR310C	YPR185W	YPR185W

[†] The 144 predictions made by MULTIPROSPECTOR are listed in three columns. In each row, there are two yeast proteins within each column, which are predicted as interacting proteins by MULTIPROSPECTOR. All interacting proteins are downloaded from MIPS.

BLAST; 1781 predictions are made by PSI-BLAST and 1215 have counterparts in DIP. The predictions having counterparts in the DIP are more likely to be true interactions than those that do not; however, because >70% of the interactions in the DIP were determined by yeast two-hybrid screening with a potentially high false-positive rate, the predictions still need to be validated by more reliable experimental methods. Further analysis of these predicted pairs will be presented in another manuscript that is currently in preparation.

Estimation of the False-Negative Rate for Yeast-Predicted Interactions

We have to remind the readers that it is incorrect to simply divide the interactions that we are unable to predict by the total number of MIPS interactions as the

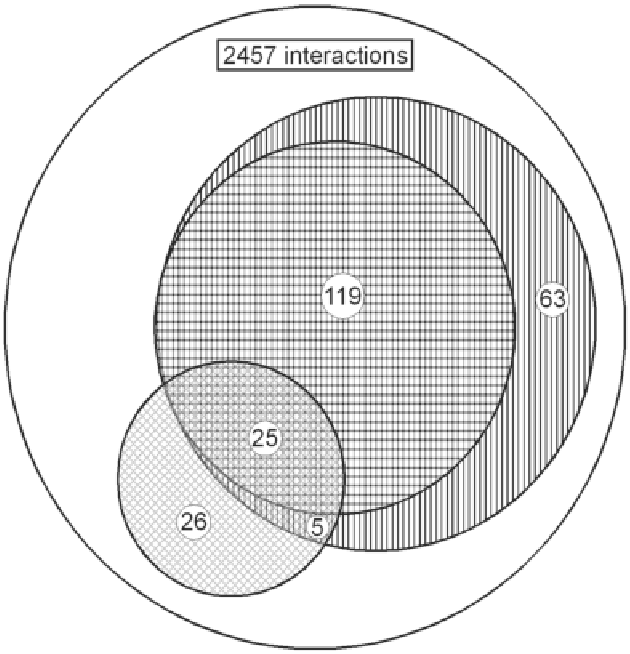

Fig. 6. Comparison of the predictions made by MULTIPROSPECTOR and a simple approach using PSI-BLAST. The biggest circle represents the entire data set of 2457 yeast interactions. Within this circle, there are three shaded circles. The vertically shaded circle represents the 212 possible interactions identified after phase I threading. The horizontally shaded smaller circle represents the 144 interactions predicted by MULTIPROSPECTOR. The lower left shaded circle represents the 56 predictions made by a simple approach using PSI-BLAST.

false-positive rate. There are two reasons. First, it is well known that yeast two-hybrid screening has a potentially high false-positive rate.[40,41] MIPS is composed of the interactions primarily determined by the yeast two-hybrid method (>70%), which leaves most of the interactions in MIPS to be validated by more reliable experimental methods. Second, it also depends on how the yeast two-hybrid screening is performed (e.g., how the bait and prey clones are constructed). This results in an overlap rate as low as 15.7% (150 of 957) by two different groups.[40,41] Our overlap rate is <15% with MIPS, which is primarily due to our limited dimer database. A similar effect was also observed in the work of other researchers.[4]

Because there is no straightforward way of obtaining the false-negative rate, here we suggest estimating the false-negative rate for the predictions of yeast-interacting proteins in the MIPS database from a functional point of view. The false-negative rate of predictions can be calculated by the ratio of interactions we missed to the interactions we should have predicted. Our approach to obtain the number of interactions we should have predicted is based on the following assumptions. First, if either monomer is strongly predicted to have a structure different from that in our dimer database, then it is excluded because this indicates that the dimer database is incomplete. Of the 2457 protein–protein interactions in MIPS, 485 interactions have at least one protein confidently predicted as monomers with a template structure by PROSPECTOR that is absent in our dimer database (Z-score ≥ 5.0) and subsequently excluded. Second, if both yeast proteins of any interacting pair in MIPS are known to have the same function as a protein complex in our dimer database, MULTIPROSPECTOR is likely to predict these interactions. This approach will provide useful hints on the actual false-negative rate; however, we are fully aware that functions and structures are not perfectly correlated. Of the remaining 1972 protein–protein interactions in MIPS, 629 interactions have at least one protein annotated as "unknown" function. For the remaining 1343 interactions, we perform key word matches with our dimer database; 244 interactions involving 266 unique proteins with both interacting partners are found having the same annotations as complexes in our dimer database. As we have shown in the previous section, 144 interactions are predicted by our approach; thus, the apparent false-negative rate is estimated to be 41%. In nearly a half of the remaining interactions that have functional counterparts in our dimer database, one of the protein partners does not have hits by the monomer threading in phase I, which means they might still be predicted by improvement of our protocols, such as introducing the new threading algorithm that is mentioned in Discussion.

Web Site Access

For non-commercial users, MULTIPROSPECTOR can be accessed on our Web site at http://bioinformatics.danforthcenter.org/services/proint/, and the resulting predictions will be e-mailed to the user, typically within 24 h. Included in these predictions are the predicted association state and the monomer structure, and when appropriate, the pair of proteins predicted to interact and a model of the predicted quaternary dimeric structure.

DISCUSSION

In the current work, we have developed MULTIPROSPECTOR, a multimeric threading protocol for recognizing protein–protein interactions and predicting the complex structures. The method has been tested against biological monomers and homodimers and has been shown to successfully recognize their oligmerization states. The test set on heterodimers showed that by using the current method, we can predict the interactions between different proteins. We have also tested the method by predictions of the interactions of yeast proteins, with initially encouraging results.

The multimeric threading algorithm, MULTIPROSPECTOR, proposed here has four advantages. First, it is better than methods based only on sequence homology, such as the simple approach described in Materials and Methods, or just single-chain threading. The basis of sequence homology-based methods and single-chain threading is that if the query sequence has the sequence identity or threading Z-score above a certain threshold (e.g., an E-value < 0.01 for a sequence-based method, and a Z-score > 5.0 for PROSPECTOR) with the template sequence, then the query sequence is considered to have the same fold. For example, suppose A and B are a pair of interacting proteins with known structures. If query se-

TABLE VI. Predicted Yeast Interactions by a Simple Approach Using PSI-BLAST[†]

Predicted pairs		Predicted pairs		Predicted pairs	
YNL331C	YNL331C	YNL030W	YNL031C	YPR165W	YDR389W
YOR128C	YOR128C	YPL204W	YLR182W	YJL026W	YGR180C
YOR128C	YBR134W	YDR171W	YDR171W	YIL074C	YIL074C
YLR109W	YLR109W	YLR309C	YLR309C	YIL074C	YER081W
YPR185W	YPR185W	YOL081W	YNL098C	YNL243W	YNL243W
YIL159W	YKR055W	YLR347C	YNL189W	YER029C	YLR147C
YFR028C	YFR028C	YJL124C	YBL026W	YFL008W	YFR031C
YLR229C	YDL135C	YJL124C	YER112W	YFL008W	YJL074C
YMR168C	YMR168C	YBL026W	YLR438C-A	YLR275W	YPR182W
YEL061C	YEL061C	YBL026W	YER146W	YNL333W	YFL059W
YCR005C	YCR005C	YBL026W	YLR275W	YJR159W	YDL246C
YGL019W	YOR039W	YLR438C-A	YER112W	YMR236W	YGL112C
YBR109C	YML057W	YER112W	YER146W	YGR144W	YGR144W
YBR109C	YLR433C	YPL174C	YBR079C	YHR025W	YHR025W
YAL003W	YPR080W	YNR032W	YGR123C	YEL021W	YEL021W
YAL003W	YBR118W	YER023W	YER023W	YBR137W	YBR137W
YBR009C	YBR010W	YER099C	YOL061W	YEL017W	YEL017W
YBR009C	YNL031C	YDL135C	YPR165W	YKL067W	YKL067W
YNL030W	YBR010W	YDL135C	YKR055W		

[†]The 56 predictions made by a simple approach using PSI-BLAST are listed in three columns. In each row, there are two yeast proteins within each column, which are predicted as interacting proteins by the simple approach using PSI-BLAST using a PSI-BLAST E-value of 1.0. All interacting proteins are downloaded from MIPS.

quences X and Y are found to be similar, respectively, with A and B, then by using a simple sequence-based or single-chain-threading method, X and Y would be predicted to interact with each other. However, the same folded structures do not necessarily have the same degree of association as shown by the examples of 1bkz and 1slt. The present multimeric threading approach takes into account the interfacial energies and thus is able to address this issue; in addition, the alignment could change to reflect interfacial interactions.

The second advantage is that although multimeric threading uses structural information, it does not require that the structures of the query proteins be solved. In this sense, it is more widely applicable than a docking approach. For example, most of the yeast proteins do not have solved structures; thus, their complex structures are unable to be assigned by using docking. However, it remains to be established just how accurate the predicted structures are; this issue will be addressed in future work.

The third advantage is that multimeric threading automatically gives the binding site. Because the binding surface is usually difficult to characterize, here we only concentrate on the binding partners that have been observed before.

The fourth advantage is that a set of empirical indicators has been developed to determine whether a complex is formed. This is unlike docking where only the best possible docked complex can be provided, with no assessment of whether in fact the complex will actually form.

In addition, multimeric threading sometimes improves the sensitivity of fold recognition in regular threading. For example, single-chain threading did not align 1bjf to the template 1alv with enough confidence (i.e., a Z-score $>$ 5.0). But by incorporating the interfacial potentials, the Z-score is improved to $>$5.0, and the fold of 1alv is recognized.

However, the relatively small number of predictions from the MIPS database predictions suggests the limitations of our method as follows.

First, the method depends on the number of solved protein complexes in the PDB. Given that on average each protein might have 2–10 binding partners,[25] the nearly 14,000 protein structures in the entire PDB (July 2001) could form 28,000–140,000 complexes. Currently, the number of complexes in PDB is about 3000, which is 2–10% of the possible number. This is the main reason why when we perform multimeric threading on the yeast-interacting protein database, only 144 of 2457 interactions have been predicted by using structures in our dimer database. As the size of the multimeric database expands, the number of predictions will increase. Currently, our multimeric database is built of only PDB records that contain two chains. The PDB records that have more than two chains have not been processed. By incorporating the multi-chain PDB records, the interface database can and will be expanded.

Second, multimeric threading also depends on the performance of the single-chain-threading method, which is the first part of multimeric threading. We have seen in several cases that if the threading fails to identify the correct fold at a medium confidence level in phase I, we will not be able to find it in the second phase of multimeric threading. At this stage, our single-chain-threading algorithm, PROSPECTOR, is able to assign ~30% of the full proteins (not just fragments of the proteins) encoded by the entire yeast genome, which is comparable with other researchers' work.[50,51] This 30% limitation is due to both the incompleteness of the fold in PDB and the threading algorithm itself.

Thus, the effectiveness of MULTIPROSPECTOR is partly bound by the capacity of the single-chain-threading algorithm. This results in a false-negative rate of 41%. Recently, an improved version of PROSPECTOR was developed (Skolnick, in preparation), which improves the yeast monomer-threading prediction to ~58%. We will test the improvement of protein–protein interaction predictions by using the new version of threading in the near future.

Third, the interfacial energy could also account for the false predictions. As we can see in Figure 6, our potentials misassigned 11% of true dimers and 15% of proteins having crystallization artifacts under current threshold. This might be improved by including distance-dependent terms or detailed atomic interactions.

The seeming discrepancy between the high-success rate of the test cases and the low-prediction outcomes from MIPS database is because our dimer database is derived from PDB and thus is not a comprehensive set of multimeric complexes templates.

Comparison of our method with PSI-BLAST suggests that our method generally outperforms the simple approach using PSI-BLAST in the test cases and in the applications to the yeast-interacting proteins. But because the prediction results of these two methods are not completely overlapped, it is possible for us to use PSI-BLAST to improve our method. However, when we incorporate PSI-BLAST results, we should be aware that the same folds do not always have the same degree of association as shown by the examples of 1bkz and 1slt. That is, the procedure cannot differentiate between monomers and dimmers. The interfacial energy screening in our algorithm should be included to validate the predictions.

Despite the limitations listed above, MULTIPROSPECTOR appears to be promising in assigning structures to protein–protein interactions. The ability to predict protein–protein interactions in the yeast genome indicates the algorithm's genomic scale applicability.

To summarize, we have proposed a straightforward way to predict protein–protein interactions using multimeric threading. An empirical standard has been established, and the test cases on yeast show that MULTIPROSPECTOR can predict a significant number of protein–protein interactions. Application to additional genomes is now underway.

ACKNOWLEDGMENTS

We thank Ms. Julie Heger for her assistance in the preparation of this article. We also acknowledge Dr. Adrian K. Arakaki's contribution in the preparation of the figures.

REFERENCES

1. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular biology of the cell, 3rd ed. New York: Garland Publishers; 1994.
2. Frieden C. Protein–protein interaction and enzymatic activity. Annu Rev Biochem 1997;40:653–696.
3. Legrain P, Wojcik J, Gauthier JM. Protein–protein interaction maps: a lead towards cellular functions. Trends Genet 2001;17: 346–352.
4. Aloy P, Russell RB Interrogating protein interaction networks through structural biology. Proc Natl Acad Sci USA 2002;99:5896–5901.
5. Aloy P, Oliva B, Querol E, Aviles FX, Russell RB. Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. Protein Sci 2002;11:1101–1116.
6. Fields S, Song O. A novel genetic system to detect protein–protein interactions. Nature 1989;340:245–246.
7. Williams NE. Immunoprecipitation procedures. Methods Cell Biol 2000;62:449–453.
8. Bollag DM. Gel-filtration chromatography. Methods Mol Biol 1994;36:1–9.
9. Hansen JC, Lebowitz J, Demeler B. Analytical ultracentrifugation of complex macromolecular systems. Biochemistry 1994;33:13155–13163.
10. Doyle M. Characterisation of binding interactions by isothermal titration microcalorimetry. Curr Opin Biotechnol 1997;8:31–35.
11. Lakey JH, Raggett EM. Measuring protein–protein interactions. Curr Opin Struct Biol 1998;8:119–123.
12. Qin J, Vinogradova O, Gronenborn AM. Protein–protein interactions probed by nuclear magnetic resonance spectroscopy. Methods Enzymol 2001;339:377–389.
13. Wuthrich K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. Science 1989;243:45–50.
14. Elcock A, Sept D, McCammon A. Computer simulation of protein–protein interactions. J Phys Chem 2001;105:1504–1518.
15. Jones S, Thornton JM. Prediction of protein–protein interaction sites using patch analysis. J Mol Biol 1997;272:133–143.
16. Bock JR, Gough DA. Predicting protein–protein interactions from primary structure. Bioinformatics 2001;17:455–460.
17. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
18. Janin J. Protein–protein recognition. Prog Biophys Mol Biol 1995;64:145–166.
19. Gilson M K, Honig B. Calculation of the total electrostatic energy of a macro-molecular system: solvation energies, binding energies, and conformational analysis. Proteins 1988;4:7–18.
20. Helmer-Citterich M, Tramontano A. PUZZLE: a new method for automated protein docking based on surface shape complementarity. J Mol Biol 1994;235:1021–1031.
21. Warwicker J. Investigating protein–protein interaction surfaces using a reduced stereochemical and electrostatic model. J Mol Biol 1989;206:381–395.
22. Jackson RM, Gabb HA, Sternberg MJ. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
23. Vakser IA. Evaluation of GRAMM low-resolution docking methodology on the hemagglutinin-antibody complex. Proteins 1997; Suppl 1:226–230.
24. Jones S, Thornton JM. Principles of protein–protein interactions. Proc Natl Acad Sci 1996;93:13–20.
25. Norel R, Petrey D, Wolfson HJ, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;36: 307–317.
26. Blattner FR, et al. The complete genome sequence of Escherichia coli K-12. Science 1997;277:1453–1474.
27. Marcotte EM, Pellegrini M, Ng H, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein–protein interaction from genome. Science 1999;285:751–753.
28. Bowie J, Luthy R, Eisenberg D. Method to identify protein sequences that fold into known three-dimensional structures. Science 1991;253:164–170.
29. Godzik A, Skolnick J, Kolinski A. A topology fingerprint approach to the inverse folding problem. J Mol Biol 1992;227:227–238.
30. Skolnick J, Kolinski A. A unified approach to the prediction of protein structure and function. Adv Chem Physics 2002;120:131–192.
31. Koretke KK, Russell RB, Copley RR, Lupas AN. Fold recognition using sequence and secondary structure information. Proteins 1999;Suppl 3:141–148.
32. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
33. Sippl MJ, Weitckus S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a

database of known protein conformations. Proteins 1992;13:258–271.

34. Panchenko A, Marchler-Bauer A, Bryant SH. Threading with explicit models for evolutionary conservation of structure and sequence. Proteins 1999;Suppl 3:133–140.
35. Altshul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
36. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
37. Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins: Structure, Function, and Genetics 2001;42:319–331.
38. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
39. Mewes HM, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C, Stocker S, Weil B. MIPS: a database for genomes and protein sequences. Nucleic Acids Res 2000;28:37–40.
40. Utez P, Rothberg JM. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 2000;403: 623–627.
41. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001;98:4569–4574.
42. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
43. Rizzi M, Wittenberg JB, Coda A, Fasano M, Ascenzi P, Bolognesi M. Structure of the sulfide-reactive hemoglobin from the clam Lucina pectinata. Crystallographic analysis at 1.5 A resolution. J Mol Biol 1994;244:86–99.
44. Libeu CA, Kukimoto M, Nishiyama M, Horinouchi S, Adman ET. Site-directed mutants of pseudoazurin: explanation of increased redox potentials from X-ray structures and from calculation of redox potential differences. Biochemistry 1997;36: 13160–13179.
45. Bolognesi M, Gatti G, Menagatti E, Guarneri M, Marquart M, Papamokos E, Huber R. Three-dimensional structure of the complex between pancreatic secretory trypsin inhibitor (Kasal type) and the trypsinogen at 1.8 A resolution. Structure solution, crystallographic refinement and preliminary structural interpretation. J Mol Biol 1982;162:839–868.
46. Gros P, Fujinaga M, Dijkstra BW, Kalk KH, Hol WG. Crystallographic refinement by incorporation of molecular dynamics: thermostable serine protease thermitase complexed with eglin c. Acta Crystallogr B 1989;45:488–499.
47. Tsien RY. The green fluorescent protein. Annu Rev Biochem 1998;67:509–544.
48. Gallet X, Charloteaux B, Thomas A, Brasseur R. A fast method to predict protein interaction sites from sequences. J Mol Biol 2000;302:917–926.
49. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 2002;30:303–305.
50. Baker DSA. Protein structure prediction and structural genomics. Science 2001;294:93–96.
51. Jones DT. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol 1999;287: 797–815.