Development of Unified Statistical Potentials Describing Protein-Protein Interactions

Hui Lu, Long Lu, and Jeffrey Skolnick
Donald Danforth Plant Science Center, St. Louis, Missouri 63132

ABSTRACT A residue-based and a heavy atom-based statistical pair potential are developed for use in assessing the strength of protein-protein interactions. To ensure the quality of the potentials, a nonredundant, high-quality dimer database is constructed. The protein complexes in this dataset are checked by a literature search to confirm that they form multimers, and the pairwise amino acid preference to interact across a protein-protein interface is analyzed and pair potentials constructed. The performance of the residue-based potentials is evaluated by using four jackknife tests and by assessing the potentials' ability to select true protein-protein interfaces from false ones. Compared to potentials developed for monomeric protein structure prediction, the interdomain potential performs much better at distinguishing protein-protein interactions. The potential developed from homodimer interfaces is almost the same as that developed from heterodimer interfaces with a correlation coefficient of 0.92. The residue-based potential is well suited for genomic scale protein interaction prediction and analysis, such as in a recently developed threading-based algorithm, MULTIPROSPECTOR. However, the more time-consuming atom-based potential performs better in identifying near-native structures from docking generated decoys.

INTRODUCTION

Knowledge of the full sequence of a genome contains explicit information about the identity of individual proteins, not with whom they interact. However, protein-protein interactions constitute a very important aspect of protein function (Valencia and Pazos, 2002). Thus, the development of tools capable of identifying such interactions is very important (Pazos et al., 1997; Hu et al., 2000; Landgraf et al., 2001; Ma et al., 2001).

Over the past decade, both experimental and computational studies of quaternary structure associated with protein-protein interactions have been a very active field (Conte et al., 1999; Glaser et al., 2001). Many insights have been gained, but the understanding of protein-protein interactions is far from complete. For example, docking, a method for the predicting of the structure of a protein-protein complex, still has only limited success even when two protein-protein partners have known experimental structures (Smith and Sternberg, 2002). There has been considerable controversy about protein-protein surface composition, residue preferences, and protein-protein interaction mechanisms (Zhang et al., 1999; Sheinerman et al., 2000; Elcock and McCammon, 2001).

A recently developed novel protein interaction prediction program, MULTIPROSPECTOR (Lu et al., 2002), has been able to generalize our threading algorithm PROSPECTOR (Skolnick and Kihara, 2001) to predict protein-protein interactions and their corresponding quaternary structures with the addition of a new protein-protein interaction potential term. The quality of the prediction from MULTIPROSPECTOR depends on the quality of the interfacial potentials used to assess the strength of protein-protein interactions. Statistical potentials have been shown to be quite successful in structure prediction on the single domain level (Bonneau and Baker, 2001; Kihara et al., 2001; also see Proteins special issue S5, 2001, on CASP4); here we want to extend this approach to the prediction of protein-protein interactions.

There have been several publications on the construction of protein-protein interaction potentials (Vajda et al., 1997; Moont et al., 1999; Glaser et al., 2001; Jiang et al., 2002). Some have been around for quite a while and are based on the relatively fewer experimental structures of protein complexes that are available (Robert and Janin, 1998; Ponstingl et al., 2000); still others have problems in defining the reference state or other shortcomings to fit our purpose. There are two reasons why we have revisited the issue of protein-protein interaction potentials. First, there are more multimeric structures now solved; second, we wish to carefully evaluate the potential with systematic test cases to ensure the quality of our potential.

To extend the approach from protein tertiary structure prediction to protein quaternary structure prediction, we built the protein-protein interaction potential following the same procedure used in the construction of the monomer potential. We also carefully checked the details as to how the potential was constructed, as done for the monomer potential (Skolnick et al., 2000). Another issue we would like to address is the difference, if any, between homodimers and heterodimers, as we would expect their interfacial interactions to be the same.

Submitted September 24, 2002, and accepted for publication October 30, 2002.

Address reprint requests to Jeffrey Skolnick, Buffalo Center of Excellence in Bioinformatics, 901 Washington St., Buffalo, NY 14203. E-mail: skolnick@buffalo.edu.

Hui Lu's present address is Dept. of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607.

Long Lu's present address is Dept. of Biochemistry and Molecular Biophysics, School of Medicine, Washington University, St Louis, MO 63110.

Jeffrey Skolnick's present address is Center of Excellence in Bioinformatics, University at Buffalo, 901 Washington St., Buffalo, NY 14203.

© 2003 by the Biophysical Society

0006-3495/03/03/1895/07 $2.00

We also develop a detailed atomic pair statistical potential. As shown previously, our atomic monomeric pair potential performs better than the residue-based pair potential for near-native structure selection (Lu and Skolnick, 2001). In a similar spirit, we build an atomic potential for protein-protein interactions (Lu and Skolnick, 2001). Although an atomic pair potential is not yet practical for genome scale interaction predictions because of the computational cost, we do explore its use in the near-native protein-complex structure selection.

The organization of this paper is as follows: In the Methods section, we first describe the construction of the dimer database, then the construction of the statistical protein-protein interaction pair potentials. In the Results section, we first present a self-consistent test to see if the potentials recognize the protein complexes from our database; then we present a test that distinguishes true dimers from false ones. We further use both the residue and atomic potentials to select near-native structures from docking generated decoys. The residue-based statistical potential also plays a key role in a multimeric threading protocol; this is briefly discussed. In the Discussion section, we summarize our approach and analyze the limitations of the current method.

METHODS

Dimer structure selection

The Protein Data Bank (PDB) is a database of solved three-dimensional protein structures (Berman et al., 2000), some of which are crystallized with more than a single protein chain. To study the statistics of protein-protein interactions and to construct a statistical pair potential, the selection of a representative set is very important. In this work, we will concentrate only on dimers. Our dimer dataset, DIMER-1, was constructed by selecting co-crystallized records from the PDB that satisfy the following criteria: First, the resolution of each chain should be less than 2.5 Å. Second, each chain in the dimer database should have more than 30 amino acids. Third, the number of dimeric interacting residue pairs is at least 30. 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. Fourth, to make sure nonredundant complexes are included in the database, we require that no pair of complexes have any chain with a sequence identity larger than 35%. Fifth, we extensively search the literature to confirm that the structures selected are experimentally validated dimers rather than crystallization artifacts. These genuine dimers are explicitly listed as ''biological dimer'' in the literature, as assessed by methods such as coimmunoprecipitation and gel filtration.

In total, 340 protein dimer structures (271 homodimers and 69 heterodimers) are selected according to the procedure of DIMER-1. We have extended DIMER-1 by the following ways: 1), we reduced the resolution criterion from 2.5 Å to 3 Å; and 2), we added the structures published in the last six months. This dataset is called DIMER-2, which consists of 768 protein complexes (617 homodimer and 151 heterodimers). DIMER-2 includes DIMER-1 and has an additional 428 protein complexes. The lists of both DIMER-1 and DIMER-2 can be found on our group's web site (http://bioinformatics.buffalo.edu/multimer).

Statistical analysis

The residue composition of protein-protein interfaces has been analyzed. Interfacial residues are defined as those having at least one heavy atom in one chain that is less than 4.5 Å away from any heavy atom in the other chain. A surface residue is defined as being more than 40% solvent-exposed for its heavy atoms, with the remainder defined as buried residues. We have used DIMER-1 as the dataset to calculate the residue composition of protein interfaces. The residue composition of the surface residues, buried residues, and all residues are calculated from a selected PDB dataset with 1191 monomeric structures that have less than 35% sequence identity between any of the two proteins in that dataset (Lu and Skolnick, 2001).

Interfacial statistical potentials used for multimeric threading

The statistical interfacial pair potentials are developed from the dimer datasets DIMER-1 and DIMER-2. The interfacial pair potentials, $P(i, j)$, are calculated by examining each interface of the protein complex using the following formula:

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

where $N_{\text{obs}}(i,j)$ is the observed number of interacting pairs 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:

$$N_{\text{exp}}(i, j) = X_{\text{i}} \times X_{\text{j}} \times N_{\text{total}}, \quad (2)$$

where X_{i} is the mole fraction of residue type i and is calculated as N_{i}/N. N_{total} is the total number of interacting pairs. There are various ways of counting N, the total number of residues, and N_{i}, the total number of residues of amino acid type i. For example, one can count all the residues in the protein, count only the surface residues, or count only the protein-protein interfacial residues. By applying Boltzmann's principle to the ratio of the observed frequencies to expected frequencies of pairings between two residue types, one obtains a statistical potential between those two residue types.

As in other derivations of a statistical pair potential, the choice of reference state (the number of expected contacts if there are no preferential interactions) is very important. From the above formula, a key factor is how to calculate the mole fraction, X_{i}. Different ways have been tested; the best results come from the calculation where only surface residues are used for the computation of the mole fraction.

A heavy atom-based distance-dependent statistical potential is also built to describe interfacial pair interactions. We use a residue-specific definition for each heavy atom type, i.e., a Cα in leucine is different from a Cα in lysine. In total, there are 167 types of heavy atoms for all the amino acids. The strategy and computation procedures are the same as those used in the preparation for the single chain atomic pair potential (Lu and Skolnick, 2001), except that here the potentials are constructed only when two atoms are from different chains.

These pairwise statistical potentials presented here can also be found on our web site.

Test cases

The test sets used in discriminating real protein-protein interfaces from artificial ones were taken from a publication by Thornton's group (Ponstingl et al., 2000). The docking-generated decoys were taken from the Vakser group's web site (http://reco3.ams.sunysb.edu/data/decoy/database.html) and the Sternberg group's web site (http://www.bmm.icnet.uk).

RESULTS

Statistical analysis of the interface

The residue composition of the protein-protein interaction interface has been calculated. Fig. 1 plots the mole fraction

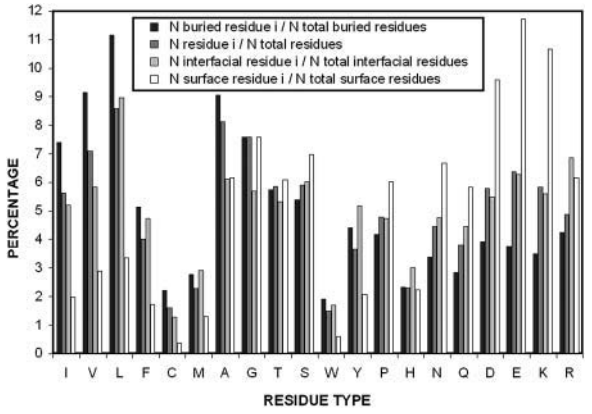

FIGURE 1 The residue composition for the whole protein, for buried residues, for surface residues, and for the protein-protein interfacial residues. The x axis lists the 20 amino acids according to their hydrophobicity and the y axis plots the percentage of each amino acid. The surface residue composition has a low percentage of hydrophobic residues and a high percentage of charged residues. The buried residue composition has a high percentage of hydrophobic residues and a low percentage of charged residues. The protein-protein interface and the whole protein compositions are similar.

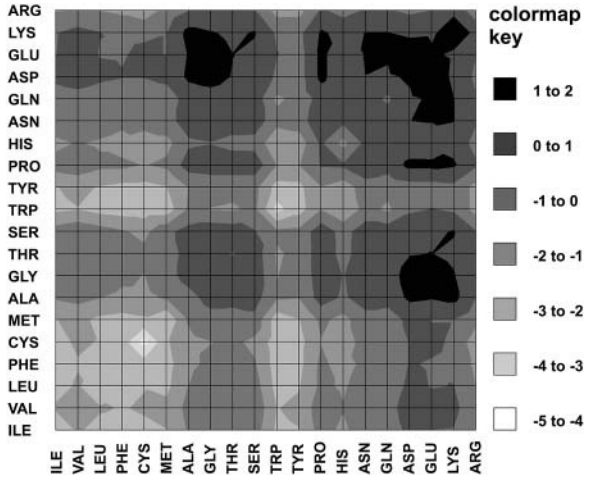

FIGURE 2 Contour representation of the unified 20×20 residue-based pairwise potential for protein-protein interactions. The numerical values of the potential can be found on our web site (http://bioinformatics.buffalo.edu/multimer).

of each residue for the overall composition, the surface residue composition, buried residue composition, and protein-protein interaction site residue composition. The figure shows that the interfacial composition is very similar to the overall composition, but is different from both the surface residue composition and the buried residue composition. In the interface region, the mole fractions of hydrophobic residues are larger than those of the surface region but less than those of the buried region. The overall composition, the surface residue composition, and the buried residue composition, respectively, are the same when calculated with DIMER-1 complexes and with the PDB-select monomer database. The observation is similar to that reached by other workers (Glaser et al., 2001).

Residue-based potential and self-consistent test

The 20×20 residue-based potential constructed from DIMER-1 is plotted in Fig. 2. Hydrophobic residues are attractive to each other and hydrophilic residues are repulsive. For example, Leu has an attractive potential with all residues except Lys, Glu, and Asp. Besides the Cys-Cys pair that may form a disulfide bond, the most attractive pairs are between hydrophobic residues pairs such as Leu, Ilu, and Phe. The most repulsive pairs are Asp-Asp, Lys-Lys, Glu-Glu, Glu-Gly, and Asp-Gly. This pattern is similar to what has been observed in the statistical potential obtained from monomers.

The first (minimal) test is whether we can successfully obtain a favorable potential energy ($E < 0$) for DIMER-1 complexes by using the potential constructed from DIMER-1. This is a necessary but not sufficient condition for a potential to be useful. For 94% of the protein complexes, the interface potential energy is favorable ($E < 0$).

The second test is whether the protein complexes in DIMER-1 satisfy the energy threshold used to distinguish between true dimer interfaces and false ones. We have determined that this threshold is -15 from an evaluation of true and false dimers (see below). Details of how this value is obtained are presented in the next subsection. In DIMER-1, 86% of the complexes satisfy the -15 threshold.

To make sure that the potential did not simply memorize the protein-protein surface composition and the interaction pairs, we performed four jackknife tests. In each case, a randomly selected set of 75% of the DIMER-1 protein complexes are used to construct the potential; the remaining 25% of the complexes are used as a test set. Inasmuch as all proteins in DIMER-1 have less than 35% sequence identity among each other, the complexes in the test set are not homologous to the complexes used in the potential construction. The correlation coefficient between the original potential with any of the jackknife potentials is 0.99. In all four tests, the fraction of test set complexes with a favorable energy ($E < 0$) range from 93% to 96%; the fraction that satisfy the threshold ($E < -15$) range from 84% to 86%.

We further used the potential constructed from DIMER-1 to evaluate complexes in the DIMER-2 dataset; the fraction of energy-favorable ($E < 0$) complexes is 89% and the ratio that satisfies the -15 energy threshold is 81%. If we use the potential constructed from DIMER-2 to evaluate the DIMER-1 and DIMER-2 datasets, the fraction of favorable complexes ($E < 0$) is 93% and 89%, respectively. The fractions that satisfy the threshold $E < -15$ are 85% and 80%, respectively. A summary of these results is presented in Table 1.

TABLE 1 Comparison of protein-protein interfacial and monomeric pair potentials

	DIMER-1 complexes		DIMER-2 complexes	
	$E < 0$	$E < -15$*	$E < 0$	$E < -15$*
DIMER-1 potential	94%	86%	89%	81%
DIMER-2 potential	93%	85%	89%	80%
Monomer potential	69%	59%†	63%	54%†

*−15 is the threshold for dimer assignment.
†For monomer potential, the threshold is −3.

As mentioned before, the selection of a reference state is very important. We have tested a number of cases using the mole fraction of the whole protein, instead of the mole fraction of surface residues, in calculating the number of expected contacts. The percentage of complexes with a favorable energy ($E < 0$) dropped to 84% from 94% and the percentage of complexes that satisfied the threshold ($E < -5$ for the potential using the whole protein composition as the reference state; $E < -15$ for the potential using surface residue composition as reference state) dropped to 68% from 86%.

Comparison of homodimer and heterodimer potentials

We have tested the difference between potentials constructed from a homodimer dataset and a heterodimer dataset. The correlation coefficient between the homodimer potential and heterodimer potential is 0.92. The comparisons of these two potentials are listed in Table 2. Using potentials developed from a homodimer dataset to evaluate a heterodimer dataset, we can have a favorable energy ($E < 0$) in 96% of the cases and a lower than threshold energy ($E < -15$) in 87% of the cases. Using a potential developed from a heterodimer dataset to evaluate a homodimer dataset, we can have $E < 0$ in 93% of the cases and $E < -15$ in 82% of the cases. Furthermore, the correlation coefficient between the DIMER-1 potential and the homodimer (heterodimer) potential is 0.99 (0.95). We conclude that there isn't much of a difference between potentials constructed from homodimers, heterodimers, or both combined. Thus, we will use the potential built from the combination of homodimers and heterodimers in all further applications.

TABLE 2 Comparison of the homodimer and heterodimer interfacial potentials

	Homodimer set		Heterodimer set	
	$E < 0$	$E < -15$*	$E < 0$	$E < -15$*
homodimer potential	94%	85%	96%	87%
heterodimer potential	92%	82%	98%	87%

*−15 is the energy threshold for dimer assignment.

Comparison of protein-protein and monomeric protein potential

Previously, our group has developed several pair potentials for monomeric protein structure prediction (Skolnick et al., 2000), which perform quite well in ab initio structure prediction (Kihara et al., 2001). Here, we compare the performance of these potentials with the newly derived dimer potential in terms of their respective ability to discriminate dimer interfaces (Table 1). We have found that in only 69% of the cases does the protein-protein interface have a favorable potential energy ($E < 0$) when evaluated with the monomer potential, as compared with 96% using the protein-protein potential. We have noticed that a different threshold should be used when checking the monomer potential. Using the same method in which the threshold of −15 is determined for protein-protein potential, the threshold for the monomer-based potential is set to be −3. With this new threshold, 59% of the complexes can be distinguished, as compared with 86% using the protein-protein potential with threshold −15. The percentage of complexes that satisfy the threshold of −15 for the monomer-based potential is only 28%. Even though the correlation between the protein-protein potential and the monomer-based potential is quite high, viz., 0.86, their respective discriminative ability is very different.

Discriminating true from artificial interfaces

The goal of the development of our protein-protein interaction potential is to predict true multimeric complexes. We have used our potential to test a published dataset (Ponstingl et al., 2000). This dataset includes one group of proteins that are true dimers and a second group of proteins that have artificial crystallization interfaces. When our residue-based potentials are used, we correctly assign 90% of the cases to be a dimer or a monomer, which is similar to the published result with various sequence- and structure-based methods (Ponstingl et al., 2000; Elcock and McCammon, 2001).

A heavy atom-based protein-protein potential is also built following the procedure in a previous work (Lu and Skolnick, 2001). Evaluation of the DIMER-1 database with this potential results in 96% of the complexes with favorable energy ($E < 0$), and 90% of the complexes have energies lower than the threshold ($E < -88$; the determination of the threshold is described below). When we use this newly constructed atomic protein-protein statistical potential to evaluate the true dimer interfaces from the false ones, the discrimination ratio increases to 95%, compared to 90% when the residue-based potential is used.

In Figs. 3 and 4, the energies of both true and false complexes are plotted. In both the residue-level and the atomic-level potential evaluation, the true protein-protein surfaces have lower energies than false ones in general. The overlap of these two groups is only 10% for residue-based

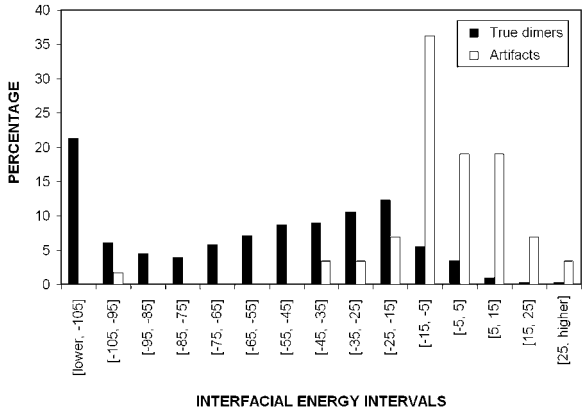

FIGURE 3 The energy distribution for true dimers and false ones evaluated with the residue-based pair potential. The energy threshold that separates these two groups the most is −15. Using this threshold, only 10% of cases are wrongly assigned. When the same procedure is applied using the monomer-based residue pair potential, the threshold is −3, and the fraction of wrongly assigned complexes increased to 41% (data not shown).

potentials and 5% for atom-based potentials. From these figures we can find a cutoff value that is most efficient in separating true protein-protein interfaces from the false ones. The cutoff value for the residue-based potential is −15, and is −88 for the atom-based potential. With the same procedure, the cutoff of the monomer pair potential is −3; however, it can only correctly assign ~60% of the cases.

Near-native structure selection from docking decoys

The atomic pairwise statistical potential is used to select near-native structures from docking generated decoy sets.

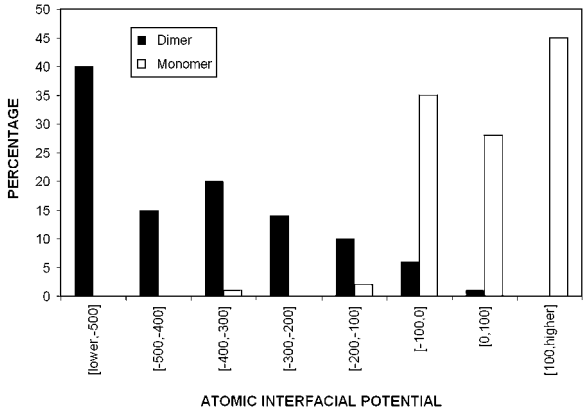

FIGURE 4 The energy distribution for true dimers and false ones evaluated with the heavy atom-based interfacial pair potential. Most true dimers have a potential energy less than −100, and most false dimers have a potential energy larger than −100. The energy threshold that separates these two groups the most is −88. With this threshold, the percentage of wrongly assigned complexes is only 5%.

We have used 16 decoy sets downloaded from the Sternberg group's web site and five decoy sets downloaded from the Vakser group's web site. In each of these decoy sets, there is one native complex structure and three near-native complex structures with 96 wrong decoys (total of 100 structures). In 20 out of the 21 test sets, the native structure ranked in the top five lowest energy selections; in 11 cases, the native structure has the lowest energy. In 15 complexes, a near-native structure can be selected in the top five lowest energy structures. We have separated these 21 complexes into two groups: 15 dimers and six trimers. The average rank of native structures for the 15 dimer decoy sets is 1.4, and the corresponding rank for the six trimer decoy sets is 4.8. In Table 3, the ranking of protein complexes' native and near-native structures, as well as the root mean-square deviations of the near-native docking decoys, are listed.

The residue-based potential was also tested with these decoy sets, and the results are listed in Table 3. The performance of this residue potential is worse than that of the atomic potential. In only 10 of the cases were the native structures selected in the top five. However, when we separately check the results of dimer and trimer decoys, in 10 of the 15 dimer cases, the native structures are selected in the top five, with three of them ranked in the top one. The average rank for the native structures in dimer decoys is 6.9. The average rank for native structures in trimer decoys is 22. It appears that our residue pair potential built on dimer interfaces does not work particularly well on trimers and presumably higher order complexes. On the other hand, the atomic potential didn't show significant difference in the performance of selecting native and near-native structures in dimer and trimer decoys.

Multimeric threading

One major application of the residue-based interface potential is as part of our multimeric threading algorithm MULTIPROSPECTOR. This program is based on a previously published threading program for single protein structure prediction (Skolnick and Kihara, 2000). The goal is to recognize potential quaternary structures on a genome scale. In the first step, each protein is assigned a possible fold; then, for those proteins whose template is part of a dimer, a second round of dynamic programming is performed and the interfacial energies are evaluated. When using the interfacial residue-based pair potential with the threshold value of −15 that was obtained in the last section, we have successfully recognized different dimeric states of proteins that have more than 50% sequence identity. A test using MULTIPROSPECTOR with this unified residue potential has been able to predict both homodimer and heterodimer interactions with 92% accuracy in a test set of 55 complexes (Lu et al., 2002). A detailed evaluation of the method is presented elsewhere (Lu et al., 2002), and currently the method has been used in the

TABLE 3 Near-native docking decoy ranking with statistical interfacial pair potential

PDB*	R (res, native)†	R (atom, native)‡	R (res, near)§	R (atom, near)¶	RMSD (A)‖
			Dimers		
1avz	24	2	38	4	5.1
1bgs	15	1	3	3	2.6
1brc	4	1	1	2	1.6
1cgi	1	1	2	4	4.8
1dfj	2	4	61	9	4.6
1fss	10	1	6	2	3.2
1ugh	1	1	5	1	6.0
1wq1	11	1	37	4	5.5
2pcc	4	1	6	6	5.0
2sic	1	1	2	2	2.1
1chg-1hpt	2	3	1	4	1.0
1sup-2ci2	5	2	1	5	1.0
2ptn-4pti	4	1	1	2	1.0
5cha-2ovo	2	1	1	7	7.7
1a2p-1a19	18	4	11	4	1.0
average	6.9	1.6	11.7	3.7	
			Trimers		
1ahw	38	3	15	4	2.0
1bvk	42	4	46	15	8.3
1dqj	17	4	45	12	15.1
1mlc	10	3	11	14	9.4
1wej	15	1	14	3	3.1
2kai	10	14	13	4	3.2
average	22.0	4.8	24.0	8.6	

*PDB code. The decoys with a single PDB code are from Sternberg's group; the decoys with two PDB codes are from Vakser's group (see text).
†The ranking of native structure using the residue-based dimer potential.
‡The ranking of native structure using the atom-based dimer potential.
§The best ranking of near-native structure using the residue-based dimer potential.
¶The best ranking of near-native structure using the atom-based dimer potential.
‖The best root mean-square deviation in the top selections using the atom-based dimer potential.

genome scale interaction prediction of yeast protein-protein interaction (Lu et al., submitted).

DISCUSSION

In the current work, we have developed statistical potentials for the evaluation of the protein-protein interactions and for threading-based quaternary structure prediction. The potential at the residue level can quickly evaluate the interfacial energy and is very useful for genome scale quaternary structure prediction. The atomic level potential has better discriminatory power, but it takes a longer time to evaluate the stability of a structure complex. Thus, it is more appropriate for near-native structure selection from docking generated decoys.

Consistent with other work (Glaser et al., 2001), a detailed analysis shows that the residue composition associated with protein-protein interfaces is similar to the overall composition of the whole protein and differs from the compositions of both buried and surface residues. This might be the reason that using a hydrophobic patch analysis for protein-protein site prediction does not always work (Conte et al., 1999; Hu et al., 2000; Elcock et al., 2001).

Even though the residue composition of the protein-protein interface and of the whole protein are very similar, the statistical potential developed for monomer structure prediction (monomer potential) is not very good in the application to protein-protein binding site prediction and evaluation. Although the correlation between the monomer and protein-protein potentials is as high as 0.86, the discriminative power of the potential depends on the details. In only 59% of the cases do protein-protein interaction interfaces have a lower energy than the threshold ($E < -3$) when evaluated with the monomer potential, as opposed to 85% when evaluated with the dimer potential ($E < -15$). Thus, a dimer potential seems necessary.

An interesting result is that the potentials constructed from the homodimers and the heterodimers have very similar performance. The correlation coefficient between these two potentials is 0.92. Furthermore, when we use the potential from a homodimer to evaluate a heterodimeric interface, the performance is very similar to that when we use the potential from a heterodimer dataset. To maintain a unified approach, we decided to use the potential constructed from the whole dimer database, including both homodimers and heterodimers.

Many protein complexes, including most homodimers, undergo two-state folding thermodynamics. In two-state binding complexes, the subunits cannot fold independently without association. These binding surfaces are generally larger and more apolar than those of three-state binding complexes such as that involving antibody-antigen complexes (Jones and Thornton, 1996). In this sense, two-state association is similar to monomer folding. However, our results show that the monomer potential performs differently (and on average worse) than the dimer potential, but is on average highly correlated to it. The potential developed from homodimers (most are two-state binding complexes) and the potential developed from heterodimers (some are three-state binding complexes) have very similar performance and can, as a practical matter, be used interchangeably. Although one might have expected a different result reflecting the possibly different composition of the interface of two-state and three-state multimers, in practice the same potential can be used on both. Perhaps, this reflects the fact that the reference state and composition effects are appropriately accounted for.

By scoring the true and false protein-protein interfaces, we can determine a threshold that would best discriminate real from false dimer interactions. This threshold is important for use in genome scale prediction by MULTIPROSPECTOR. We were able to correctly assign in 90% of the threading dimer test cases that include true and false dimers, partly due to the good discriminatory ability of this potential (Lu et al., 2002). We have noticed that some proteins with similar

structures and high sequence identity may have different degrees of association; one may be a dimer (1slt), whereas the other closely related one is a monomer (1bkz) (Lu et al., 2002). The potential has been proved to be successful in selecting dimers from monomers even when the sequence identity between them is more than 50%.

The application of statistical potentials in docking decoy selection is of great value. It is a more difficult task than selecting threading candidates. It is worth mentioning that the performance on dimer decoy set selections is better than the trimer decoys. The atomic potential clearly outperforms the residue potential; in most cases the native and near-native structures are the top five lowest energy complexes.

We gratefully thank Ms. Julie Heger for her assistance in the preparation of this article.

This research was supported in part by National Institutes of Health grants GM-48835 and GM RR-12255.

REFERENCES

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Bonneau, R., and D. Baker. 2001. Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* 30:173–189.

Conte, L., C. Chothia, and J. Janin. 1999. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285:2177–2198.

Glaser, F., D. Steinberg, I. Vakser, and N. Ben-Tal. 2001. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins.* 43:89–102.

Elcock, A., and A. McCammon. 2001. Identification of protein oligomerization states analysis of interface conservation. *Proc. Natl. Acad. Sci. USA.* 98:2990–2994.

Elcock, A., D. Sept, and A. McCammon. 2001. Computer simulation of protein-protein interactions. *J. Phys. Chem. B.* 105:1504–1518.

Hu, Z., B. Ma, H. Wolfson, and R. Nussinov. 2000. Conservation of polar residues as hot spots at protein interfaces. *Proteins.* 39:331–342.

Jiang, L., Y. Gao, F. Mao, Z. Liu, and L. Lai. 2002. Potential of mean force for protein-protein interaction studies. *Proteins.* 46:190–196.

Jones, S., and J. Thornton. 1996. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA.* 93:13–20.

Kihara, D., H. Lu, A. Kolinski, and J. Skolnick. 2001. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA.* 98:10125–10130.

Landgraf, R., I. Xenarios, and D. Eisenberg. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* 307:1487–1502.

Lu, L., H. Lu, and J. Skolnick. 2002. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by muttimeric threading. *Proteins.* 49:350–364.

Lu, H., and J. Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins.* 44:223–232.

Ma, B., H. Wolfson, and R. Nossinov. 2001. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. *Curr. Opin. Struct. Biol.* 11:364–369.

Moont, G., H. Gabb, and M. Sternberg. 1999. Use of pair potentials across protein interfaces in screening predicted docked complexes. 35:364–373.

Pazos, F., M. Helmer-Citterich, G. Ausiello, and A. Valencia. 1997. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* 271:511–523.

Ponstingl, H., K. Henrick, and J. M. Thornton. 2000. Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins.* 41:47–57.

Robert, C., and J. Janin. 1998. A soft, mean-field potential derived from crystal contacts for predicting protein-protein interactions. *J. Mol. Biol.* 283:1037–1047.

Sheinerman, F., R. Norel, and B. Honig. 2000. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol.* 10:153–159.

Skolnick, J., and D. Kihara. 2001. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins.* 42:319–331.

Skolnick, J., A. Kolinski, and A. Ortiz. 2000. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins.* 38:3–16.

Smith, G., and M. Sternberg. 2002. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* 12:28–35.

Vajda, S., M. Sippl, and J. Novotny. 1997. Empirical potentials and functions for protein folding and binding. *Curr. Opin. Struct. Biol.* 7:222–238.

Valencia, A., and F. Pazos. 2002. Computational methods of the prediction of protein interactions. *Curr. Opin. Struct. Biol.* 12:368–373.

Zhang, C., J. Chen, and C. DeLisi. 1999. Protein-protein recognition: exploring the energy funnels near the binding sites. *Proteins.* 34:255–267.