PROTEINS: Structure, Function, and Genetics 46:190–196 (2002)

Potential of Mean Force for Protein–Protein Interaction Studies

Lin Jiang,[1,2] Ying Gao,[1,2] Fenglou Mao,[1,2] Zhijie Liu,[1,2] and Luhua Lai[1,2]*

[1]*Institute of Physical Chemistry, College of Chemistry and Molecular Engineering, Peking University, Beijing, China*
[2]*State Key Laboratory for Structural Chemistry Studies of Stable and Unstable Species, Beijing, China*

ABSTRACT **Calculating protein–protein interaction energies is crucial for understanding protein–protein associations. On the basis of the methodology of mean-field potential, we have developed an empirical approach to estimate binding free energy for protein–protein interactions. This knowledge-based approach has been used to derive distance-dependent free energies of protein complexes from a nonredundant training set in the Protein Data Bank (PDB), with a careful treatment of homology. We calculate atom pair potentials for 16 pair interactions, which can reflect the importance of hydrophobic interactions and specific hydrogen-bonding interactions. The derived potentials for hydrogen-bonding interactions show a valley of favorable interactions at a distance of ≈3 Å, corresponding to that of an established hydrogen bond. For the test set of 28 protein complexes, the calculated energies have a correlation coefficient of 0.75 compared with experimental binding free energies. The performance of the method in ranking the binding energies of different protein–protein complexes shows that the energy estimation can be applied to value binding free energies for protein–protein associations. Proteins 2002;46:190–196.**

© 2001 Wiley-Liss, Inc.

Key words: pair potentials; knowledge-based potentials; protein association; protein–protein interaction; protein recognition

INTRODUCTION

Understanding principles of protein recognition that are pertinent to biological process is one of the long-term goals in the area of protein science. Theoretical approach to estimate the binding affinity of protein–protein associations plays an important role in protein function and design studies. There has been increasingly interest in scoring the energies of protein–protein complexes and participating in protein–protein docking contests (e.g., Strynadka et al.[1]).

Several empirical approaches have been developed to predict binding affinity and to discriminate between correct and incorrect protein–protein interactions by running molecular dynamics or Monte Carlo (MC) simulations. Practical empirical approaches can be classified into three groups, according to the different free energy scoring functions selected. The first group includes force field-based methods that predict binding free energies using a "master" thermodynamic equation,[2] including a free energy perturbation (FEP) method.[3] In these methods, the success of estimating relative binding free energies of protein complexes has been shown,[4,5] and the native protein conformation can be distinguished from decoy conformations.[6] However, the computational burden is still prohibitive, and these methods are not fast enough to screen a large structural database against a given target. The second group, regression methods,[7,8] adopt a linear regression approach to describe binding energies, in which the free energy is written as the sum of important physically meaningful contributions, such as rotational and translational entropy loss, hydrophobic and hydrophilic surface areas, the number of frozen torsions, and those parameters are fitted to experimental binding data. The methods have relatively accurate predictions when applied within the model range. But it is not clear to what extent that they can be applied to complexes not included in the training set. Thus, these methods may not be transferable among different systems. Moreover, available experimental data of binding free energy are far too few to enable the task, in respect to protein–protein interactions.

The third group is the statistical method, which includes mainly the mean field potentials (PMF) that have appeared to be successful in protein folding studies.[9–12] In this approach, the structural frequencies observed from a known structural training database are converted into contact or pairwise potentials. Many applications have shown its usefulness in protein-ligand binding studies, a subset of the protein recognition process. Wallqvist et al.[13] assessed the binding affinities of HIV-1 protease inhibitor by using a function based on atom–atom surface propensities from a series of 38 protein–ligand complexes. DeWitte and his coworkers[14,15] derived a coarse-grained model (SMOG) and applied it to the de novo design of new compounds. Recently, two approaches were successfully applied to estimate the binding affinity of different protein–

Grant sponsor: Ministry of Science and Technology of China; Grant sponsor: Natural Science Foundation of China; Grant number: 29525306; Grant sponsor: Committee of Science and Technology of Beijing.

*Correspondence to: Luhua Lai, Institute of Physical Chemistry, College of Chemistry and Molecular Engineering, Peking University, Beijing, China 1000871. E-mail: lhlai@pku.edu.cn or lai@mdl.ipc.pku.edu.cn

Received 3 May 2001; Accepted 6 September 2001

© 2001 WILEY-LISS, INC.
DOI 10.1002/prot.10031

ligand complexes and give good correlation with experiments: Bleep (Mitchell et al.[16,17]) was derived from a representative training set containing a broad variety of protein–ligand structures, with a careful treatment of homology. Muegge and Martin[18] presented the simplified potential approach that treated solvation and entropic contributions implicitly, with the definition of an appropriate reference state and the introduction of a correction term accounting for the volume taken by the ligand.

By way of contrast, the nature of protein–protein interfaces is neither similar to the interior of the individual protein monomers nor to that of protein–ligand interactions. Thus, it is inappropriate to apply to protein–protein associations the existing mean field energy functions that are derived from protein–ligand complexes directly. Moont et al.[19] used empirical residue–residue pair potentials to screen possible complexes for protein–protein dockings. Because of the small size of the data sets available, the pair potentials generated from protein–protein complexes did not give good result. Robert and Janin[20] derived a series of novel PMF for predicting protein–protein interactions, in which a number-based parameterization is introduced. The coarse potential treated just two atom types: hydrophobic and hydrophilic, which is similar to hydrophobic-polar (HP) model. This representation obviously improved statistical reliability on the residue-based formulation of Moont.[19] And they used a simple on-off atomic contact model, depending on whether the distances between atoms are within 6.05 Å; hence, detailed interface geometry is generally lacking. Some statistical reports[21,22] have shown that there is a significant population of charged and polar residues at protein–protein interfaces, which tend to reflect the composition of protein surfaces rather than protein interiors.

Hydrophobic residues tend to scatter over the entire surface and form small patches that are interspersed with charged and polar residues, which shows that protein–protein and protein–ligand interaction are significantly different. The coarse potential only reflects hydrophobic interactions, not including the specificity of hydrogen bond interactions.

Although the simple model appears to be successful in folding simulations, it may be inappropriate when applied to complex protein–protein associations. Robert and Janin's potentials[20] have been applied to discriminate the correct protein–protein complex structure from decoy conformations but have not been applied to score the binding energies of different protein–protein complexes.

An empirical approach for predicting protein–protein interaction energies is developed here. We have derived a distance-based PMF by using the nonredundant training set containing a wide range of structures with the treatment of homology. Four atom types were defined for this purpose: hydrogen bond donor, hydrogen bond acceptor, both donor and acceptor, and neutral type (neither donor nor acceptor). Based on the simple and appropriate definition of atom types, the atom-level potentials can reflect not only the hydrophobic interactions but also the hydrogen bond interactions at protein interfaces. With the implicit treatment of solvation and entropic effects, the binding affinity of protein complexes is directly estimated without any fitting procedure used by other knowledge-based methods. Like other mean-field potentials, the empirical approach can calculate fast enough to estimate the free energies of millions of decoy conformations produced by docking programs. When it is applied to score different protein–protein complexes for binding free energies, a good performance is shown compared with experiments. We expect that the empirical energy estimation can find wide applications, especially in docking studies.

MATERIALS AND METHODS

Data Set of Known Protein–Protein Complex Structures

The data are from the archived files (updated in April 2000) of protein complex structures in the Protein Data Bank (PDB).[23] We focus on protein complexes, rather than homo-oligomeric and hetero-oligomeric entries. The molecular information provided in the PDB header files was checked. This yields 696 entries.

Except for antigen–antibody complexes, high-resolution X-ray structures of protein complexes are limited to the resolution of better than 2.5 Å. And a threshold of 3.0 Å is chosen for antigen–antibody complexes. NMR structures and theoretical models are not included.

Protein homology should be treated carefully to generate a nonredundant training data set. The redundant homologous entries need to be eliminated. We have calculated the homology scores according to the methodology of Myers and Miller.[24] First, PDB entries are dissected into "receptor protein–ligand protein" pairs, which are non-covalently bound and contain no less than five normal amino acids in both parts. We also rely on the molecular information provided in the PDB header file. If both parts of the two pairs have >70% sequence similarity, the pairs are considered as homology pairs, and the lower resolution representative is eliminated. If there is >70% sequence similarity between one part of the two pairs, and 30–70% between the other part of the two pairs, a subjective evaluation has to be made to judge the similarity of the interface, with the aid of the interactive computer graphics software of QUANTA (Molecular Simulation Inc., San Diego, CA). The filtered set includes 191 "receptor protein–ligand protein" pairs from 179 PDB entries (Table I), containing antigen–antibody complexes, enzyme–protein inhibitor complexes, and protease–peptide complexes in a wide range of data sets.

Definition of Atom Types

We define four atom types: hydrogen bond donor, hydrogen bond acceptor, both donor and acceptor, and neutral atom (neither donor nor acceptor). The specific definition of four types is shown in Table II. In our method, each of 16 protein–protein atom pairs contact through either hydrogen bond or hydrophobic interactions, which are ubiquitous on protein–protein interfaces and tend to reflect the forces driving the association of protein complexes. Each protein complex is dissected into "receptor protein–ligand

TABLE I. The 179 PDB Entries of the Data Set

1aly	1a2k	1a2x	1a2y	1a3r	1a4y	1a94	1acb	1ak4	1apm
1aqd	1ava	1avg	1avw	1axi	1ay7	1aya	1azs	1azz	1b0n
1b2s	1b33	1bai	1bbz	1bc5	1be9	1bgx	1bii	1bj1	1bjr
1blx	1bog	1bt6	1bth	1bvn	1bxi	1c3q	1ca0	1ca9	1ce1
1cgi	1cho	1cjf	1cjr	1cka	1cm1	1cn3	1cxz	1d2z	1d3b
1d4t	1d4v	1dan	1dhk	1dkx	1dvf	1eai	1eay	1eer	1efn
1efu	1ekb	1evh	1f58	1fak	1fbi	1fc2	1fdl	1fle	1flt
1gc1	1ggi	1gua	1gux	1hia	1hlt	1iai	1iak	1ibr	1igc
1ikn	1ir3	1itb	1jhl	1jrh	1jxp	1kig	1lcj	1lck	1ldt
1lfd	1lgb	1lpb	1mct	1mda	1mhc	1mkx	1mpa	1ncb	1nfd
1nmc	1noc	1nrs	1nsg	1nsn	1oak	1osp	1osz	1pdk	1pip
1pyt	1qav	1qc6	1qfu	1qja	1qle	1qmz	1qo0	1qsn	1qur
1rsu	1sbb	1sbw	1sgp	1sha	1shd	1slu	1sm3	1smp	1smr
1spb	1spp	1srn	1stf	1str	1tbr	1tco	1tgs	1tmc	1tmq
1tx4	1tze	1ucy	1ugh	1uug	1vad	1vpp	1vrk	1vwg	1wej
1wq1	1www	1x11	1yag	1ycp	1ycq	1ycs	1zfp	2ap2	2cbl
2clr	2cwg	2igf	2jel	2mip	2mta	2pcc	2prg	2seb	2sic
2tgp	2trc	3erd	3hfm	3pro	4er4	4htc	4sgb	5apr	

TABLE II. Definition of Protein Atom Types

Type	Description	Definition
A	Hydrogen bond acceptor	Backbone O, OD of ASN, OE of GLN, ND of HIE, and NE of HID
D	Hydrogen bond donor	NE of TPR, ND of ASN, NE of GLN, NE of HIE, ND of HID, and N of LYS and ARG
N	Neutral atom (neither acceptor nor donor)	Carbon and sulfur atoms, N of PRO
B	Both acceptor and donor	Hydroxyl oxygen, ND NE of HIS and carboxyl oxygen in C-terminal

protein" pair. The receptor protein part and ligand protein part are treated differently. Generally the receptor protein part is the bigger part in a complex, such as the antibody protein in an antigen–antibody complex, the enzyme in an enzyme–protein inhibitor complex, or the protease in a protease–peptide complex. The atoms of the same type between receptor protein part and ligand protein part should be differentiated. So we calculate 16 atom pair interaction terms here. The atom occupancy in the crystal structure file is used to function as a weighting coefficient.

In our training set, metal and small ligand molecules are excluded. And it is assumed that all crystallized complexes using water as the medium. Water molecules are neglected, because the solvation effects are implicitly treated. And hydrogen atom types are omitted in all the analysis, because most of the complexes in the PDB contain no hydrogen atoms.

Calculation of PMFs and Estimation of Binding Energy

Pair potentials are derived from our training set according to the methodology of Sippl.[9] According to reverse Boltzmann relationship, the protein–protein interaction free energy between the receptor-protein atom of type i and the ligand-protein atom of type j at a distance r can be written as

$$A_{ij} = -kT \ln[f_{ij}(r)/Z_{ij}] \quad (1)$$

where k is the Boltzmann constant and T is the absolute temperature; $f_{ij}(r)$ is a frequency of these ij contacts occurring at distance r.

In fact, the statistical potential we derived is the difference of the potential of the atom pair versus that of a reference potential,

$$A_{ij} - \text{reference energy} \equiv \Delta A_{ij}(r) = kT \ln[1 + m_{ij}\sigma] - kT \ln\left[1 + m_{ij}\sigma \frac{g_{ij}(r)}{f(r)}\right] \quad (2)$$

where m_{ij} is the total number of contacts between types i and j, $g_{ij}(r)$ is the distribution of these contacts occurring at distance r, and $f(r)$ is the distribution of all contacts for all types at distance r. The atom pair distance r uses a histogram-based representation, and r here refers to a given bin of width 0.2 Å. σ is used as a weighting function. Hence, an atom pair with no data would be represented by the average potential for atom-type pairs. On the other hand, an atom pair with a large distribution would be represented entirely by its atom data. Here we set σ to 0.02.

Distance distributions are based on all the atom–atom distances for those pairs of four atom types. A distance cutoff value of 8 Å is set, and the atom–atom distance of >8 Å is considered as no interactions and excluded from the statistical distribution of the atom pairs involved. The observed distance distributions of the two atom types involved are stored by using bins of width 0.2 Å. In the calculation of distributions, we always use the atom occupancy as a weighting function.

The choice of reference energy is simplified: we just import a big value in very short distance (i.e., where our statistics are not included) to capture strong van der

Waals repulsive potentials in this distance range. And in the distance range of our statistics, we set our choice of the reference energy that can counteract solvation effects. Therefore, the difference for the PMF versus the reference energy (ΔAij) accounts for the energetic effect of desolvation. This choice of reference state has the simple interpretation that those contacts that are observed in the database more frequently than average is favored and vice versa. By the choice of the appropriate reference energy, we implicitly treat solvation and entropic effects and directly estimate total free energies of protein–protein complexes without any knowledge of binding affinities and fitting procedures. Thus, we can use the difference ΔAij directly to estimate binding energy containing solvation effects.

The derived potentials for the interaction of atom type i and atom type j are summed up to evaluate the total PMF value.

$$A = \sum_{ij} \Delta A_{ij}(r) \times \Delta_{ij} \tag{3}$$

$$\begin{cases} \Delta_{ij} = 0 & \text{for } r_{ij} > r_{\text{cut-off}} \\ \Delta_{ij} = p_i \times p_j & \text{for } r_{ij} \le r_{\text{cut-off}} \end{cases}$$

where r_{cutoff} is the cutoff distance of atom type pair interactions and is set to 8.0 Å and p_{i} is a weighting coefficient from the atomic occupancy. The introduction of the weighting coefficient can efficiently improve the accuracy of our potentials.

Because the PMF reflects Helmholtz free energies, the entropic contributions have been captured implicitly. We also capture other contributions to total free energy by treating them implicitly. Because our reference energy is not necessarily including all the solvation effects, it is necessary to rectify these deviations. We define a scaling factor like that of Muegge and Martin.[18] Finally, the calculated PMF value is correlated to the total binding free energy by the following equation:

$$\Delta G_{\text{binding}} = \varepsilon^* A \tag{4}$$

where a scaling factor ϵ stands for all the different terms that are treated implicitly.

RESULT

Specific Interactions of Atom Pairs

The occurrence of the atom pair within a distance of 8.0 Å is recorded. If the total occurrence of all atom pairs in the shell of a distance $r \pm \Delta r$ is <50, the contributions of all atom pairs at the distance interval are ignored because of lacking statistically sufficient data. In fact, the potentials between all atom pairs occurring in the range of 2.6–8.0 Å are derived.

Figure 1 shows the calculated potentials of mean force. The potentials for the hydrogen bond acceptor–donor interactions (AD and DA) are shown in Figure 1(a). The two-letter code refers to the atom pair interaction types, where the first letter refers to the receptor protein and the second letter refers to the ligand protein. Both potentials show a minimum at a distance of ≈3 Å, corresponding to an established hydrogen bond and then become weakly repulsive around 4.0 Å. The potentials for other similar interaction types, such as AB, BA, DB, and BA, have a very similar shape to that of AD and DA (data not shown). Figure 1(b) shows that the potential for the nonpolar interaction between neutral types (NN) is repulsive at all distances up to 4.0 Å, and slightly attractive between 4.5 and 7.0 Å. Figure 1(c) shows the potential for type A and type N (AN and NA). It is surprising that a valley of favorable interactions is observed around 3.5 Å. The potentials for other similar interaction types, such as BN and NB, have a very similar valley near the same distance range.

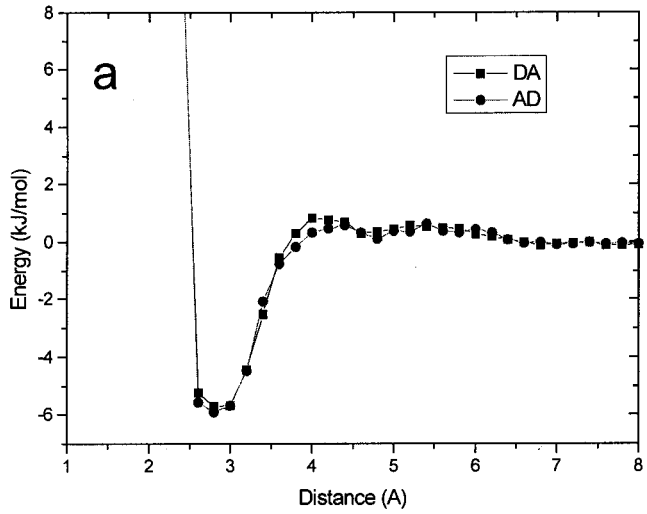
In all the cases, the occurrence of observations is sufficiently large (from 20,917 observations to 355,826 observations for every pair potential). It should be confident that the details of potentials are meaningful. All these potentials are close to zero at the cutoff value of 8.0 Å. In fact, the potentials become close to zero from about 7 Å. We also test a cutoff value of 12.0 Å, but the derived potentials remain unchanged.

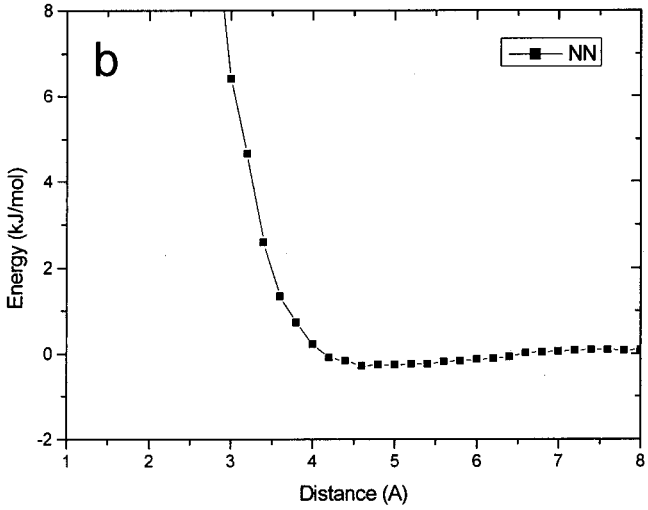
The shapes of pair potentials of AD and DA are encouraging. They are very similar to those of Sippl[25] and Mitchell,[16] which show a minimum just below 3.0 Å, with an energy barrier at about 4.0 Å. The pair potential of hydrophobic interaction (NN) also has a reasonable shape. Those are corresponding to the potential characterization of our interaction types.

Scoring Protein–Protein Complexes

We apply the scoring function of Eq. 3 to estimate a test set of 28 different protein complexes. The 28 complexes are chosen from literature by the same criterion when generating the training set of the PMF: protein complexes having resolution of 2.5 Å or better (antigen–antibody complexes are 3.0 Å) is chosen. NMR structures and theoretical models are excluded. The PDB entries and sources of the test set are shown in Table III. There are antigen–antibody complexes, protease–protein inhibitor complexes, and other complexes in our test set. We calculate PMF energies for these 28 complexes, and then the calculated values are compared with the experimental binding free energies and a correlation coefficient is calculated. We use this kind of comparison as the principal method of evaluating the success of our method for a wide range of protein–protein complexes.

By linear fitting through zero point, the calculated PMF energies give a correlation coefficient of 0.75 with experimental data. The result is presented in Figure 2. The complexes are illustrated in the figure by the PDB code (or HIVdb code). Because the short-range repulsive region of the potential is not sampled by the data, the large value of the atom pair interaction energy within 2.4 Å is imported arbitrarily as punishment. To value the effect of the arbitrary value, we test different values of the punishment score, and all the results of correlation coefficient remain 0.75. It is possible that strong van der Waals clashes in very short distance of atom pairs are scarce in protein complex crystal structures, because of the refinement of





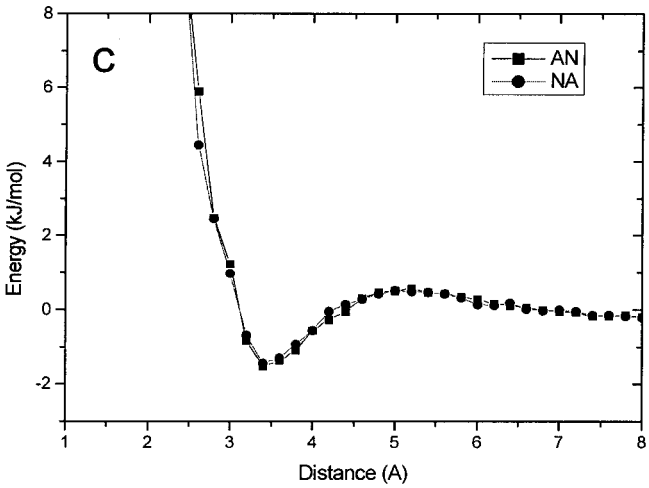


Fig. 1. The potentials of mean force. The potentials of mean force $\Delta A_{ij}(r)$ calculated with Eq. 2 are shown. The two-letter code ij refers to the atom pair interaction types, where the first letter i refers to the receptor protein and the second letter j refers to the ligand protein. Potentials are obtained at the interval of 0.2 Å, starting from 2.4 Å, and connected by smooth curves for visualization. **a:** The pair potentials for the interactions of AD and DA. **b:** The pair potentials for the interaction of NN. **c:** The pair potentials for the interactions of AN and NA.

TABLE III. The Test Set of 28 Protein Complexes[†]

1abi	1cse	1fle	1tec	2tpi	3hfl	4ins
1axi	1dkz	1igc	1tpa	2sec	3hfm	4sgb
1cgi	1dvf	1mel	1vfb	2sni	3sgb	4tpi
1cho	1fdl	1nmb	2ptc	2tgp	4htc	7hvp

[†]The experimental data of 1cse, 1tec, 1tpa, 2ptc, 2sec, 2tgp, 2tpi, and 4ins are used by Horton and Lewis.[26] The protein-protein complexes are chosen, and some protein-ligand complexes in which ligands are peptides are also chosen to test the wide-ranging ability of our method. If the same PDB entry has different data, the latest literature is chosen. The source of other binding free energies are: 1abi,[27] 1axi,[28] 1cgi,[29] 1cho,[30] 1dkz,[5] 1dvf,[31] 1fdl,[32] 1fle,[33] 1igc,[34] 1mel,[35] 1nmb,[36] 1vfb,[37] 2sni,[38] 3hfl,[39] 3hfm,[40] 3sgb,[4] 4htc,[41] 4sgb,[42] and 4tpi and 7hvp.[13] All PDB entries chosen have an appropriate resolution according to previous criterion. The structures of complexes are obtained from the Brookhaven Protein Data Bank.[23]

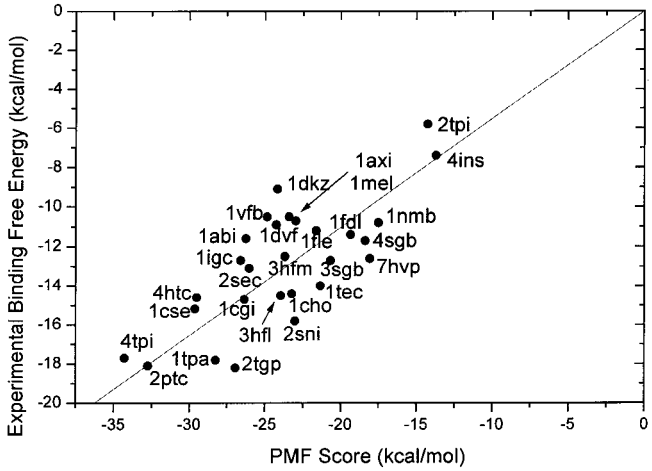


Fig. 2. Calculated PMF scores and the experimental binding free energies. A set of 28 complexes is shown. The experimental data are taken from the reference of Table III. The correlation coefficient is 0.75. The scaling factor ϵ calculated by using Eq. 4 is 1.8.

crystal structure. So the choice of the punishment score does not affect the energy calculation of the scoring crystal structures.

DISCUSSION

The atom types for protein–protein complexes used here are simpler than that of protein–ligand complexes. Because of the limited size and the more relaxed criteria in the generation of the training set, explicit definition of atom types seems to be inappropriate. Biological interfaces of protein–protein complexes contain many specific interactions, including hydrogen-bonding and water-bridging interactions. On the basis of the statistical analysis results,[22] we believe that an adequate description of hydrogen bond interactions would be necessary for protein recognition. It is obvious that this reduced representation can efficiently improve the reliability of statistics. We choose the definition of four atom types that can reflect specific hydrogen bonding interactions, rather than only hydrophobic interactions. And the implicit definition of atom types makes our potentials less sensitive, which make them more appropriate for protein docking studies. Further-

more, it is convenient to apply the potentials to value the different contribution and role of each pair interaction. Thus, the potentials may be helpful for the description of protein–protein interfaces. The potential for atom pair interaction has a reasonable shape, and the derived potentials have a good performance in the test set. These show that the simple definition of atom types can reflect the characteristics of interactions at protein–protein interfaces.

For the presence of atoms, the atom occupancy value in crystal structure is used to function as a weighting coefficient. This neglects those especially flexible atoms of side chains in crystal structures and can deal with the multiple conformations of side chains. Using the weighting function in atom presentations can improve the reliability of training data set statistics. From another point of view, the quality of training data set is improved.

When we are looking at the potential of DA atom pair interaction, a valley of favorable interaction is found at exactly the desired distance, and a reasonable fluctuation at medium and long range. That finding is similar to that of Sippl[25] and Mitchell,[16] which are, respectively, derived from peptide hydrogen bonds in proteins and protein–ligand complexes. It shows that hydrogen bond interactions have common characterizations either in proteins or in protein complexes, and they only have some difference in details. Twenty-eight protein complexes are used as the test set. It is clear that comparing the energies of different complexes is more difficult than ranking different decoy conformations of the same complex. The performance of our potentials is encouraging.

Therefore, to generate a usable potential energy, it is necessary to calculate the reference energy of average interactions. Unfortunately, this is a pragmatic task. Generally, there were two strategies for this problem: one is to convert the overall data into an energy-like term according to Boltzmann relationship and to use a suitable volume correction (e.g., Muegge and Martin[18]). The other is to import an empirical reasonable reference state (e.g., Mitchell et al.[16]). Both strategies only obtain an approximation and their methods are pragmatic. Besides those methods, another view is that the reference energy should be chosen as a set of adjustable parameters because PMF is just an empirical method. Although choosing the reference energy is theoretically pragmatic, such combined potentials based on it have succeeded in some applications. Furthermore, the potential of Mitchell[16] excluding the reference potential has the same correlation coefficient with that including the reference potential. Thus, it appears that the choice of the reference energy is not essential to our method.

All of the water molecules presented are not considered in our statistics. This means that atoms that are, in reality, exposed to the solvent may appear in the data set to be "observed" in empty space. So our PMFs have already included solvation effects to some extent. Furthermore, because a pairwise interaction occurring at the surface of the protein complexes or inside them is not distinguished in our method, the effect of the solvent is only taken into account in an average sense. In our reference state, the specificity of each contact is lost, and the remaining energetic contribution is simply from the fact that desolvation factor is replaced in very short distance where van der Waals potentials is dominant.

We treat solvation and entropic effects implicitly and choose the appropriate reference state tactfully. So the explicit consideration of energy terms is circumvented, and the total protein–protein binding energy is directly estimated. But the choice of our reference energy is quite coarse. The reference energy mainly affects the calculation of atom pair interactions in very short distance. Because there are very few strong van der Waals clashes between atom pairs in the crystal structures due to the refinement of crystal structures, the energy estimation for crystal structures will be not affected by the choice of the reference energy. The reference energy in short distance becomes very important when PMF scoring function is applied to find the right conformation in docking studies. However, it does not seem to be essential to score the energies of crystal complexes. So an explicit choice of the reference energy is not adopted in our analysis.

There are other problematic points of the definition of atom types. For instance, salt-bridge interactions are considered as hydrogen-bonding interactions, and their strong contributions to protein associations are not emphasized. Moreover, crystal interfaces are not entirely the same as biologically relevant protein–protein interfaces. Because of the high salt concentrations and the near neutral pH used in many processes of crystallization, strong salt bridge interactions have been deemphasized in statistics. When we use the data of crystal structures as the training set of the potential, the importance of charged interactions may be reduced. We might have derived PMF by combining the different atom types to a reasonable measure, and multivariate fitting methods might have been used just as in empirical scoring functions. However, our goal here is to derive a scoring function solely built on structural information. Because the available data are not enough, the fitting method and more complex atom-type definition is not adopted in our method.

CONCLUSION

We believe that the number of high-resolution protein–protein complex structures in the PDB is now sufficient to derive a PMF of general applicability in protein–protein binding studies. Taking account of homology, we derive an empirical energy scoring function from a representative training data set, without any knowledge of experimental binding affinities or any fitting procedures. Based on the simple and appropriate definition of atom types, the atom-level potentials can reflect the importance of hydrophobic interactions and hydrogen bond interactions. The potential for atom pair interaction has a reasonable shape, and the potentials of DA atom pair interaction (donor–acceptor) have a valley of favorable interaction at exactly the desired distance range, which are similar to those of Sippl[25] and Mitchell.[16] When it is applied to score 28 different protein–protein complexes for binding free ener-

gies, a correlation coefficient of 0.75 is obtained compared with experimental data. We expect that the potential of mean force developed here would be helpful to understand the interactions of protein interfaces and the forces that drive protein–protein associations.

REFERENCES

1. Strynadka NCJ, Eisenstein M, Latchaiski-Katzi E, Shcichet BK, Kuntz ID, Abagyan R, Sternberg M, James MNG. Molecular docking programs successfully predict the binding of beta-lactamase inhibitory protein to TEM-1 beta lactamase. Nat Struct Biol 1996;3:233–239.
2. Ajay, Murcko MA. Computational methods to predict binding free energy in ligand-receptor complexes. J Med Chem 1995;38:4953–4967.
3. Kollman P. Free energy calculations—applications to chemical and biochemical phenomena. Chem Rev 1993;93:2395–2417.
4. Krystek S, Stouch T, Novonty J. Affinity and specificity of serine endopeptidase–protein inhibitor interactions. J Mol Biol 1993;234: 661–679.
5. Kasper P, Christen P. Empirical calculation of the relative free energies of peptide binding to molecular chaperone Dnak. Proteins 2000;40:185–192.
6. Weng Z, Vajda S, Delisi C. Prediction of protein complexes using empirical free energy functions. Protein Sci 1996;5:614–626.
7. Böhm HJ. The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure. J Comput Aided Mol Des 1994;8:243–256.
8. Weber PC, Wendoloski JJ, Pantoliana MW, Salemme FR. Crystallographic and thermodynamic comparison of natural and synthetic ligands bound to streptavidin. J Am Chem Soc 1992;114: 3197–3200.
9. Sippl MJ. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
10. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
11. Jernigan RL, Bahar I. Structure-derived potentials and protein simulations. Curr Opin Struct Biol 1996;6:195–209.
12. Vajda S, Sippl M, Novotny J. Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 1997;7:222–228.
13. Wallqvist A, Jernigan RL, Covell DG. A preference-based free-energy parameterization of enzyme-inhibitor binding: applications to HIV-1 protease inhibitor design. Protein Sci 1995;4:1881–1903.
14. DeWitte RS, Shakhnovich EI. SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. J Am Chem Soc 1996;118: 11733–11744.
15. DeWitte RS, Shakhnovich EI. SMOG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Case studies in molecular design. J Am Chem Soc 1997;119:4608–4617.
16. Mitchell JBO, Laskowski RA, Alex A, Thornton JM. BLEEP—potential of mean force describing protein–ligand interactions. I. Generating potential. J Comput Chem 1999;20:1165–1176.
17. Mitchell JBO, Laskowski RA, Alex A, Forster MJ, Thornton JM. BLEEP—potential of mean force describing protein–ligand interactions. II. Calculation of binding energies and comparison with experimental data. J Comput Chem 1999;20:1177–1185.
18. Muegge I, Martin YC. A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 1999;42:791–804.
19. Moont G, Gabb HA, Sternberg MJE. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
20. Robert CH, Janin J. A soft, mean-field potential derived from crystal contacts for predicting protein–protein interactions. J Mol Biol 1998;283:1037–1047.
21. Xu D, Tsai C, Nussinov R. Hydrogen bonds and salt bridges across protein–protein interfaces. Protein Sci 1997;10:999–1012.
22. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
23. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80:319–324.
24. Myers EW, Miller W. Approximate matching of regular expressions. Bull Math Boil 1989;51:5–37.
25. Sippl MJ. Helmholtz free energy of peptide hydrogen bonds in proteins. J Mol Biol 1996;260:644–648.
26. Horton N, Lewis M. Calculation of the free energy of association for protein complexes. Protein Sci 1992;1:169–181.
27. Qiu X, Padmanabhan KP, Carperos VE, Tulinsky A, Kline T, Maraganore JM, Fenton JW. Structure of the hirulog 3-thrombin complex and nature of the S' subsites of substrates and inhibitors. Biochemistry 1992;31:11689–11697.
28. Atwell S, Ultsch M, Vos AMD, Wells JA. Structural plasticity in a remodeled protein–protein interfaces. Science 1997;278:1125–1128.
29. Hecht HJ, Szardenings M, Collins J, Schomburg D. Three-dimensional structure of the complexes between bovine chymotrypsinogen *A and two recombinant variants of human pancreatic secretory trypsin inhibitor (Kazal-type). J Mol Biol 1991;220:711–722.
30. Bigler TL, Lu W, Park SJ, Tashiro M, Wieczorek M, Wynm R, Laskowski M. Binding of amino acid side chains to preformed cavities: interaction of serine proteinases with turkey ovomucoid third domains with coded and coded P1 residues. Protein Sci 1993;2:786–799.
31. Olson MA, Reinke LT. Modeling implicit reorganization in continuum descriptions of protein–protein interactions. Proteins 2000; 38:115–119.
32. Novotny J. Protein antigenicity: a thermodynamic approach. Mol Immumol 1991;28:201–207.
33. Tsunemi M, Kato H, Nishiuchi Y, Kumagaye S, Sakakibara S. Synthesis and structure-activity relationships of elafin, an elastase-specific inhibitor. Biochem Biophys Res Commun 1992;185:967–973.
34. Sjobring U, Bjorck L, Kastern W. Streptococcal protein G gene structure and protein bind properties. J Mol Chem 1991;266:399–405.
35. Desmyter A, Transue TR, Ghahroudi MA, Thi MHD, Wyns L. Crystal structure of a camel single-domain VH antibody fragment in complex with lysozyme. Nat Struct Biol 1996;3:803–811.
36. Tulip WR, Harley VR, Webster RG, Novotny J. N9 neuraminidase complexes with antibodies NC41 and NC10: empirical free energy calculations capture specificity trends observed with mutant binding data. Biochemistry 1994;33:7986–7997.
37. Verhoeyen M, Milstein C, Winter G. Reshaping human antibodies: grafting an antilysozyme activity. Science 1988;239:1534–1536.
38. Svendsen I, Jonassen I, Hejgaard J, Boisen S. Amino acid sequence homology between a serine protease inhibitor from barely hordeum valgare cultivar hiproly and potato inhibitor I. Carlsberg Res Commun 1980;45:389–395.
39. Lavoie TB, Drohan WN, Smith-Gill SJ. Experimental analysis by site-directed mutagenesis of somatic mutation effects on affinity and fin specificity in antibodies specific for lysozyme. J Immunol 1992;148:503–513.
40. Padlan EA, Silverton EW, Sheriff S, Cohen GH, Smith-Gill SJ, Davies DR. Structure of the Hy/Hel-10 Fab-lysozyme complex. Proc Natl Acad Sci USA 1989;86:5938–5942.
41. Bode W, Mayr I, Baumann U, Huber R, Stone SR, Hofsteenge J. The refined 1.9 angstroms crystal structure of human alpha-thrombin: interaction with D-PHE-PRO-ARG chloromethylketone and signficance of the TYR-PRO-PRO-TRP insertion segment. EMBO J 1989;8:3467–3475.
42. Hass GM, Tidor B. Carboxypeptidase inhibitor from potatoes. Methods Enzymol 1994;80:779–790.