PROTEINS: Structure, Function, and Bioinformatics 56:93–101 (2004)

A Physical Reference State Unifies the Structure-Derived Potential of Mean Force for Protein Folding and Binding

Song Liu,[†] Chi Zhang,[†] Hongyi Zhou, and Yaoqi Zhou*
Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology and Biophysics, State University of New York at Buffalo, Buffalo, New York

ABSTRACT **Extracting knowledge-based statistical potential from known structures of proteins is proved to be a simple, effective method to obtain an approximate free-energy function. However, the different compositions of amino acid residues at the core, the surface, and the binding interface of proteins prohibited the establishment of a unified statistical potential for folding and binding despite the fact that the physical basis of the interaction (water-mediated interaction between amino acids) is the same. Recently, a physical state of ideal gas, rather than a statistically averaged state, has been used as the reference state for extracting the net interaction energy between amino acid residues of monomeric proteins. Here, we find that this monomer-based potential is more accurate than an existing all-atom knowledge-based potential trained with interfacial structures of dimers in distinguishing native complex structures from docking decoys (100% success rate vs. 52% in 21 dimer/trimer decoy sets). It is also more accurate than a recently developed semiphysical empirical free-energy functional enhanced by an orientation-dependent hydrogen-bonding potential in distinguishing native state from Rosetta docking decoys (94% success rate vs. 74% in 31 antibody–antigen and other complexes based on *Z* score). In addition, the monomer potential achieved a 93% success rate in distinguishing true dimeric interfaces from artificial crystal interfaces. More importantly, without additional parameters, the potential provides an accurate prediction of binding free energy of protein–peptide and protein–protein complexes (a correlation coefficient of 0.87 and a root-mean-square deviation of 1.76 kcal/mol with 69 experimental data points). This work marks a significant step toward a unified knowledge-based potential that quantitatively captures the common physical principle underlying folding and binding. A Web server for academic users, established for the prediction of binding free energy and the energy evaluation of the protein–protein complexes, may be found at http://theory.med.buffalo.edu. Proteins 2004;56:93–101.** © 2004 Wiley-Liss, Inc.

Key words: potential of mean force; knowledge-based potential; energy score functions; reference state; binding stability; docking decoys

INTRODUCTION

Understanding the mechanism of how proteins fold and bind is one of the central problems in molecular biology. This requires an accurate potential of mean force (or free-energy score function) that describes the water-mediated interaction between amino acid residues. One approach for obtaining the potential is using the laws of physics. However, limited by available computing power, such physical-based potentials (e.g., Brooks et al.,[1] Kaminski et al.[2]) have to be decomposed into multiple classical energy terms, each of which involves many empirical parameters. Moreover, the contribution of entropy is often not included except for effective potential for entropy of solvation.[3] An alternative but simpler approach is extracting the potential of mean force from known protein structures.[4] The resulting potentials are commonly referred to as statistical potentials, because statistical averages and distributions are used in derivation. Physical principles, however, are often violated or ignored in this statistical procedure, because the outcome depends on how statistics are made and what database is used. For example, the distance-dependent pair potential based on the commonly used Sippl approximation[5] leads to unphysical long-range repulsion between hydrophobic residues as a result of the composition difference between the core and surface of proteins.[6] The unphysical nature of the statistical potentials has limited their accuracy.

Protein folding and binding involve the same physical interactions between water and amino acid residues. It is a common practice to apply the same physical- (or semiphysical-) based potential to study them.[7–10] In contrast, significantly different compositions at the surface, core, and interface of proteins[11–13] lead to quantitatively different, distance-dependent pair potentials for folding and binding studies.[11,14] For example, Lu et al.[11] found that the success rate for discriminating a dimer interface from

[†]These two authors contribute equally to this work.

Grant sponsor: NIH; Grant numbers: R01 GM 966049 and R01 GM 068530. Grant sponsor: HHMI (to SUNY Buffalo) and the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo.

*Correspondence to: Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics and Department of Physiology and Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. Email: yqzhou@buffalo.edu

Received 24 June 2003; Accepted 2 October 2003

Published online 16 April 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20019

artificial interfaces is reduced from 86% to 59% after the database used to extract the residue-level statistical potential is changed from dimers' interfaces to monomers' structures. As a result, most existing knowledge-based potentials developed for protein–protein recognition are specific for the recognition problem by using the interfacial structures of protein complexes as the training database.[11,12,15–18] This raises an interesting question: Does a unified structure-derived potential of mean force exist for folding and binding? It is difficult to imagine that a potential trained with monomer structures can be used to study protein–protein interface, which has a different amino acid composition, unless the potential captures the essential physics behind the different "mixture" of amino acids.

The accuracy of a structure-derived potential is determined by how a reference state is defined. The most commonly used reference state for a distance-dependent pair potential is based on a composition-averaged state.[5,19,20] A reference state is required for obtaining the net interaction between amino acid residues by removing the contribution to the observed number of atomic pairs by the expected number of pairs in the absence of any interactions. Because a system of noninteracting particles is the ideal-gas state in fundamental statistical mechanics, the ideal-gas state is more likely to represent a zero-interaction state than a composition-averaged state. This forms the basis for a recently-proposed distance-scaled, finite, ideal-gas reference (DFIRE) state for deriving the distance-dependent potential of mean-force for single-chain (monomeric) proteins.[21] It was shown that the DFIRE-based potential of mean force makes a significant improvement over other knowledge-based potentials in native structure selections from decoys, prediction of mutation-induced change in folding stability, and partition of hydrophobic and hydrophilic residues.[22]

In the DFIRE state, each protein in the structural database is replaced by uniformly distributed points in a finite sphere with a diameter close to the size of the protein. A natural extension of this reference state for protein–protein complexes is to replace the entire protein complex by a finite sphere with uniformly distributed points. In contrast, for the knowledge-based potentials that use a statistically averaged state as the reference state, either the entire complex or any portion of the complex can be used for statistical averages. For example, a statistically averaged reference state can be dedicated to interfacial region of protein–protein complexes.[12,14,15,23] However, it is difficult to imagine how to apply the ideal-gas reference state only to the irregular-shaped region of interface. The use of the entire complex as the reference state makes the potential extracted from the structures of complexes essentially the same as that from monomer structures, because the statistics within single-chain proteins are dominant over those of the small interfacial region between the proteins. In other words, the physical basis of the newly developed DFIRE potential suggests that the potential should be mostly independent of structural databases (monomers or dimers) used for extraction and thus is applicable to both folding and binding.

This hypothesis is tested by the direct application of the potential previously derived from the structures of single-chain proteins[21] (termed "monomer" potential hereafter) to the protein–protein recognition problem. Specifically, the monomer potential is employed to recognize native and near-native protein–protein complexes from docking decoys and to distinguish true dimer interfaces from artificial interfaces of monomeric proteins in the crystalline state. A more critical test of the monomer potential is the prediction of the binding free energies of 69 protein complexes. The performance of this monomer potential is outstanding in all test cases. The result suggests that a physical reference state leads to a structure-derived potential that reveals the common physical interaction masked under different amino acid preferences within a single protein[22] and between protein–protein complexes.

METHODS

DFIRE-Based Potential

The derivation of equations, the method for extracting the DFIRE-based potential using a structure database, as well as the resulting potential have been described or obtained previously.[21] Here, we give a brief summary for completeness.

The atom–atom potential of mean force $\bar{u}(i,j,r)$ between atom types i and j that are distance r apart is given by[21]

$$\bar{u}(i,j,r) = \begin{cases} -\eta RT \ln \dfrac{N_{\rm obs}(i,j,r)}{\left(\dfrac{r}{r_{\rm cut}}\right)^{\alpha}\left(\dfrac{\Delta r}{\Delta r_{\rm cut}}\right) N_{\rm obs}(i,j,r_{\rm cut})}, & r < r_{\rm cut}, \\ 0, & r \geq r_{\rm cut}, \end{cases} \quad (1)$$

where $\eta = 0.0157$, R is the gas constant, $T = 300$ K, $\alpha = 1.61$, $N_{obs}(i,j,r)$ is the number of (i,j) pairs within the distance shell r observed in a given structure database, $r_{cut} = 14.5$ Å, and $\Delta r(\Delta r_{cut})$ is the bin width at $r(r_{cut})$. ($\Delta r = 2$ Å for $r < 2$ Å; $\Delta r = 0.5$ Å for 2 Å $< r <$ 8 Å; $\Delta r = 1$ Å for 8 Å $< r <$ 15 Å.) The prefactor η was determined so that the regression slope between the predicted and experimentally measured changes of stability due to mutation (895 data points) is equal to 1.0. Residue-specific atomic types were used (167 atomic types).[19,20] The number of observed atomic (i,j) pairs with the distance shell r $[N_{obs}(i,j,r)]$ was obtained from a structural database of 1011 nonhomologous (less than 30% homology) proteins with resolution <2 Å, which was collected by Hobohm et al.[24] (http://chaos.fccc.edu/research/labs/dunbrack/culledpdb.html). This database provides sufficient statistics for most distance bins (except near the hard-core regions). The average number of observed atomic pairs per bin is 655. The sufficiency of statistics is also reflected from the fact that the results for structural discrimination are insensitive to the size of structural database used to generate the potential.[21]

The exponent α for the distance dependence was obtained from the distance dependence for the number of

TABLE I. Ranking of the Native State and the *Z* Score for the 21 Docking Decoy Sets

PDB ID[a]	1chg/1hpt	1sup/2ci2	2ptn/4pti	5cha/2ovo	1a2p/1a19	1avz	1bgs	1brc
LLS[b]	3	2	1	1	4	2	1	1
DFIRE[c]	1/7.74	1/4.08	1/5.08	1/1.59	1/4.01	1/3.31	1/4.56	1/3.50
	1fss	1ugh	1wql	2pcc	2sic	1cgi	1dfj	%Success[d]
	1	1	1	1	1	1	4	10/15 (67%)
	1/4.33	1/4.85	1/5.29	1/2.86	1/4.29	1/5.89	1/4.11	15/15 (100%)
PDB ID[e]	1ahw	1bvk	1dqj	1mlc	1wej	2kai	%Success[d]	
LLS[b]	3	4	4	3	1	14	1/6 (17%)	
DFIRE[c]	1/3.80	1/3.50	1/4.85	1/4.04	1/3.10	1/4.25	6/6 (100%)	

[a]Dimers.
[b]The residue-specific all-atom knowledge-based potential due to Lu, Lu, and Skolnick derived from the interfacial structures of a dimer database.[11] The number in each cell indicates the rank of the native structure. (The *Z* score was not reported.[11])
[c]The DFIRE-based potential derived from a structure database of single-chain proteins.[21] The two numbers in each cell represent the rank of native structure and the *Z* score, respectively.
[d]The overall success rate based on the first rank.
[e]Trimers.

pairs of ideal gas points in finite spheres (finite ideal gas reference state). There are 1011 such finite spheres; each corresponds to a protein used in the structural database.[24] The radius of each sphere is cR_g and the sphere contains evenly distributed $_{nhv}$ points. Here, c is a to-be-determined constant, R_g and $_{nhv}$ are the radius of gyration and the number of heavy atoms of the corresponding protein, respectively. Constant c is determined by the number of atomic pairs in a noninteracting uniform system. The latter can be calculated from the number of atomic pairs in 1011 protein structures in the cutoff distance shell of 14–15 Å, because at that distance, we assumed zero interactions between atoms. There are about 57 million atomic pairs for 1011 proteins. c is found to be 1.225 by setting the number of atomic pairs in 1011 spheres in the 14–15 Å distance shell to 57 million. The number of pairs as a function of spatial separation, $N(r)$, is then obtained from the evenly distributed points in the 1011 spheres. The exponent α is determined by locating the optimized value α that would fit $N(r)$ with r^{α}. The optimized α value is found to be 1.61. The performance for structure discrimination using this optimized value is indeed the best compared to other α values.[21] This supports the physical procedure behind the derivation of the α value. It is not clear if this fractional exponent has any physical meaning associated with fractional exponents in packing and energy transfer studies of proteins, and if it is related to factional dimension in mathematics. For an infinite system, $\alpha = 2$ exactly.

Binding Free Energy

The total atom–atom potential of mean force, G, for each structure is given by

$$G = \frac{1}{2}\sum_{i,j} \bar{u}(_{i,j},r_{ij}) , \qquad (2)$$

where the summation is over atomic pairs that are not in the same residue and a factor of $\frac{1}{2}$ is used to avoid double-counting of residue–residue and atom–atom interactions. The binding free energy of a dimer AB is obtained as follows:

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_A + G_B) . \qquad (3)$$

Since the structures of A and B are approximated as rigid bodies and the residues at the interface contribute most to ΔG_{bind}, Eq. (3) can be further simplified to

$$\Delta G_{\text{bind}} = \frac{1}{2}\sum_{ij}^{\text{interface}} \bar{u}(_{i,j},r_{ij}) , \qquad (4)$$

where the summation is over any two atoms belonging to an "interacting" residue pair from different chains at the interface. We follow the definition of Lu et al.,[11] in which an interacting residue pair is a pair of residues from different chains that have at least one pair of heavy atoms within 4.5 Å of each other. To quantitatively test the DFIRE monomer potential, experimental binding free energies of 69 protein–protein (peptide) complexes with known three-dimensional (3D) structures were collected from the literature. To our knowledge, this is the largest database collected so far.

Structure Selections from Docking Decoys/Artificial Interfaces

The binding free energy $\Delta G_{\text{bind}}^{\text{decoy}}$ is calculated for each docking decoy (or artificial interface). The native state is correctly identified if $\Delta G_{\text{bind}}^{\text{native}}$ is the lowest value among all $\Delta G_{\text{bind}}^{\text{decoy}}$ values (the first rank). A native binding Z score (Native) is defined as $(\langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle - \Delta G_{\text{bind}}^{\text{native}})/\sqrt{\langle (\Delta G_{\text{bind}}^{\text{decoy}})^2 \rangle - \langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle^2}$, where $\langle\rangle$ denotes the average over all decoy structures of a given protein. The native Z score is a measure of the free-energy bias toward the native complex structure. To evaluate the ability to recognize near-native structures for a potential and facilitate comparison, we use the near-native Z score defined by Kortemme et al.[10]:

$$Z_{\text{score}}(\text{near native}) = \frac{\langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle_{\text{hi}} - \langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle_{\text{lo}}}{\sigma_{\text{hi}}} , \qquad (5)$$

TABLE II. The Native *Z* Scores and Near-Native *Z* Scores for the 31 Rosetta Docking Decoys

PDB ID	1a2y[a]	1qfu[a]	1cz8[a]	1wej[a]	1daj[a]	1e6j[a]	1egj[a]	1eo8[a]
KMB[b]	2.47/1.28[c]	0.01/2.15	6.04/1.66	0.79/0.28	5.80/1.50	5.28/1.76	0.72/1.42	0.96/0.48
DFIRE[d]	2.22/0.19[e]	2.45/3.12	3.15/1.14	1.25/0.70	3.05/2.15	2.12/1.75	2.04/0.66	3.69/1.42
PDB ID	1fdl[a]	1fj1[a]	1g7h[a]	1ic4[a]	1jhl[a]	1jrh[a]	1mlc[a]	1nca[a]
KMB[b]	2.66/1.20	1.51/2.58	3.38/1.99	5.29/1.97	2.31/−0.11	8.56/1.81	2.33/1.45	0.50/1.39
DFIRE[d]	2.13/0.51	2.72/2.33	2.31/0.61	2.62/1.84	1.45/−0.20	3.38/1.24	2.25/2.75	3.01/2.60
PDB ID	1nsn[a]	1osp[a]	1acb	1avz	1brs	1cho	1ugh	2pcc
KMB[b]	−0.36/0.14	7.82/0.31	11.33/2.14	1.05/0.24	3.43/2.53	12.06/3.39	2.34/1.48	−0.87/0.55
DFIRE[d]	1.04/1.08	3.26/0.59	2.40/1.09	1.16/0.27	2.50/1.35	3.10/1.17	2.40/2.20	0.77/0.66
PDB ID	1mda	1ppf	1spb	2ptc	1cse	1fin	2btf	%Success[f]
KMB[b]	−1.03/0.27	8.77/10.61	14.06/5.29	6.18/3.23	9.16/3.32	3.65/0.29	4.18/1.79	74%/71%
DFIRE[d]	0.55/1.55	2.81/1.04	4.41/2.54	2.57/1.34	3.49/1.73	5.06/1.86	2.99/1.34	94%/71%

[a]The antibody–antigen complex.
[b]The Kortemme–Morozov–Baker empirical free-energy function enhanced by orientation-dependent hydrogen-bonding potential.[10]
[c]The *Z* score (Native)/*Z* score (Near Native) from the KMB method.
[d]The DFIRE-based potential derived from a database of single-chain proteins.[21]
[e]The *Z* score (Native)/*Z* score (Near Native) from DFIRE potential.
[f]The success rate based on the number of complexes with *Z* score(Native) or Z score(Near Native) greater than 1 as in the KMB method.[10]

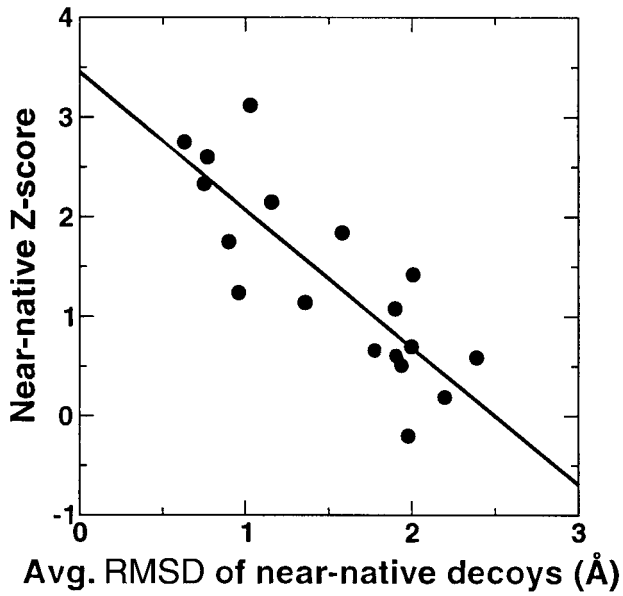

Fig. 1. The correlation between *Z* scores for near-native structures and the quality of near-native structures (the average RMSD). The solid line is the regression line, with a correlation coefficient of −0.82. This indicates that an accurate structure prediction is directly associated with the accuracy of near-native structures.

where $\langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle_{\text{lo}} (\langle \Delta G_{\text{bind}}^{\text{decoy}} \rangle_{\text{hi}})$ is the average energy score of the low (high) root-mean-square deviation (RMSD) decoys and σ_{hi} is the standard deviation of high-RMSD decoys. A decoy is considered a low-RMSD decoy if it is in the lowest 5% of RMSD distribution.[10] The low-RMSD decoys represent the near-native structures.

RESULTS

Structure Selections from Docking Decoys

The first decoy set consists of 16 and 5 decoy sets downloaded from the Sternberg group's website (http://www.bmm.icnet.uk) and the Vakser group's website (http://reco3.ams.sunysb.edu/data/decoy/database.html), respectively. The 21 decoy sets contain 15 dimers and 6 trimers. Each decoy set has 1 native complex structure and 99 decoys.

Table I compares the results of the DFIRE potential with the residue-specific, all-atom knowledge-based, Lu–Lu–Skolnick (LLS) potential.[11] The success rates of the LLS potential, which are better than earlier attempts,[25] are 10/15 (67%) and 1/6 (17%) for dimers and trimers, respectively. A significantly poorer performance for the trimer decoys suggests that the LLS potential generated from interfacial structures of dimers has a strong database dependence. In contrast, the DFIRE-based monomer potential successfully selects all 21 structures with a *Z* score of 1.59 or greater. The significantly different performance by these two potentials at the same atomic details clearly illustrates the importance of an appropriate reference state for knowledge-based potentials.

The second docking decoy set[10] (Rosetta docking decoys) contains 18 antibody–antigen complexes and 13 complexes of enzyme–enzyme and other interfaces. The list of 31 protein complexes is given in Table II. There are 400 decoys for each structure. Table II compares the *Z* scores from DFIRE-based monomer potential and those from a semiphysical, empirical free energy functional enhanced by an orientation-dependent hydrogen-bonding potential (KMB potential).[10] Kortemme et al. defined a discrimination as being successful if a *Z* score (Native) is greater than 1.0. Thus, using the same definition, the overall success rate is 23/31 (74%) and 29/31 (94%) for KMB and DFIRE potentials, respectively. DFIRE makes a significant 23% improvement. Interestingly, the ranges of *Z* scores (Native) of the two methods are very different. The *Z* score (Native) for KMB ranges from −1.03 to 14.06, whereas the *Z* score (Native) for the DFIRE potential ranges from 0.55 to 5.06. This indicates that the performance of the DFIRE potential is more stable than that of the KMB potential in native discrimination. However, in terms of discrimination between near-native decoys and other decoys, DFIRE and KMB are comparable in terms of *Z* score (Near Native). According to Kortemme et al., a near-native

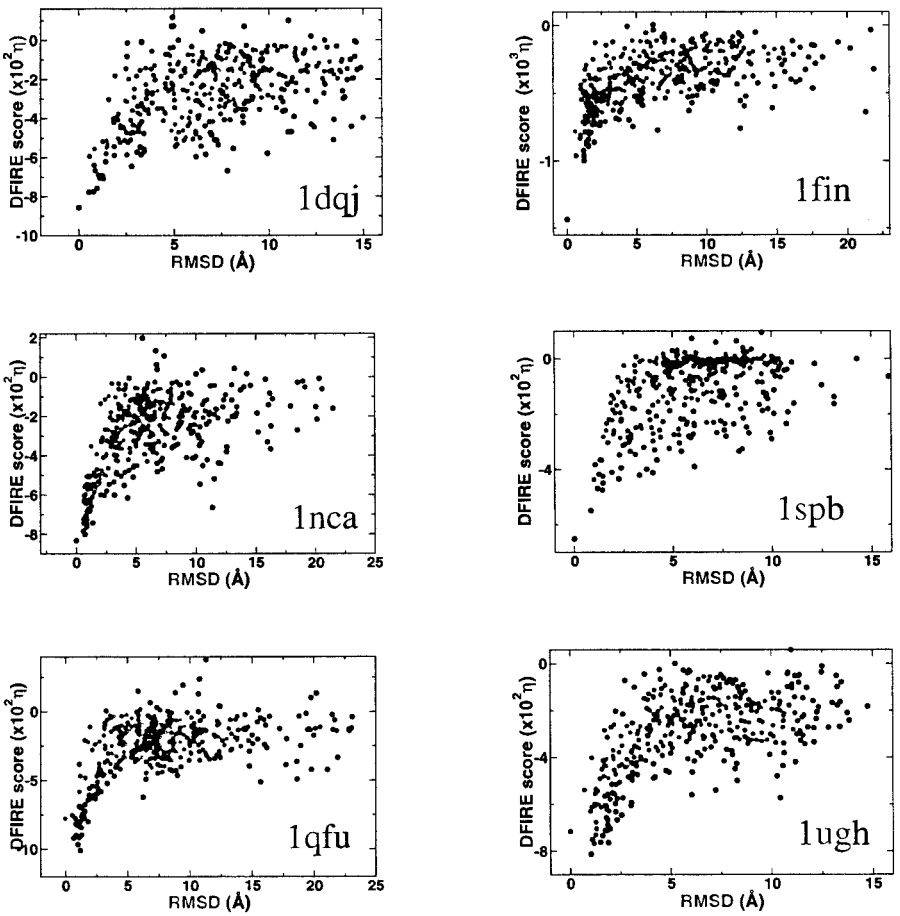

Fig. 2. Scatter plots of the DFIRE score versus RMSD of decoy from the native structure (based on C_α). Results of three proteins (1dqj, **top left**; 1nca, **middle left**; and 1qfu, **bottom left**) from the antibody–antigen complexes and three proteins (1fin, **top right**; 1spb, **middle right**; and 1ugh, **bottom right**) from the nonantibody complexes are shown.

selection is successful if a Z score (Near Native) is greater than 1.0. Both achieved a success rate of 71%.

It is surprising to find that a significant improvement in native-structure discrimination did not lead to any improvement in near-native structure discrimination. We found that the success rates of native and near-native discrimination for 18 antibody–antigen complexes are 67% and 72% for the KMB method, respectively, whereas the corresponding numbers are 100% and 61% for the DFIRE monomer potential, respectively. For 13 other complexes, the success rate of native and near-native discrimination are 85% and 69% for the KMB method, respectively, whereas the corresponding numbers are 85% and 85% for the DFIRE monomer potential, respectively. Thus, the DFIRE potential for near-native discrimination is worse than the KMB potential only for the antibody–antigen complexes, where a dramatic reduction from 100% in the native discrimination to 61% in near-native discrimination is observed. We found that there is a significant correlation between Z score (Near Native) and the quality of near-native decoys for the 18 antibody–antigen complexes. (The quality of near-native decoys is defined by the average RMSD of the near-native decoys; see Methods section). The correlation coefficient is -0.82, as shown in Figure 1. This suggests that a decoy set with better near-native decoys will improve the performance of DFIRE potential in near-native discrimination.

Another way to characterize the ability to detect near-native conformations is the correlation between energy score and RMSD, when RMSD is smaller than about 3 Å.[10] In the docking decoy set, the number of proteins whose correlation coefficients are equal to or greater than 0.5 is 18 for KMB and 23 for DFIRE. Examples are given in Figure 2 for both antibody–antigen and nonantibody complexes. Thus, the DFIRE energy function is potentially better in its ability to detect near-native conformation.

Interface Selections

The data set of 172 interfaces was established by Ponstingl et al.[17] It contains 96 monomeric crystal interfaces and 76 homodimeric interfaces. The challenge is to distinguish the true homodimeric interfaces from artificial interfaces in crystalline state. In Figure 3, the distributions of energies of both true and artificial interfaces calculated with the DFIRE potential are shown. In general, the energies of true interfaces are lower than those of

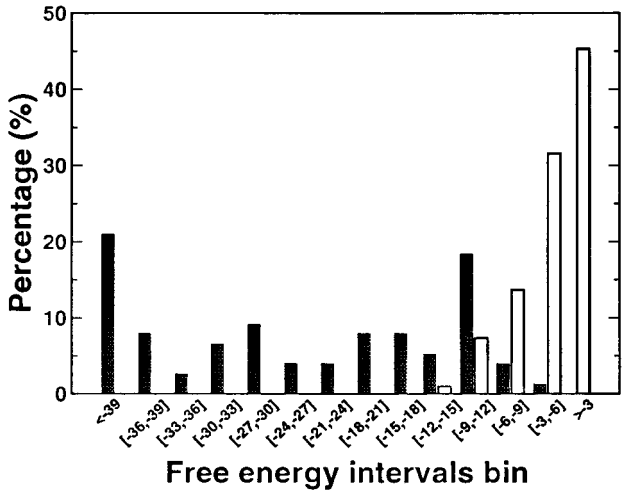

Fig. 3. The distribution of the energy scores of the artificial (open bars) and true (black bars) dimeric interface.

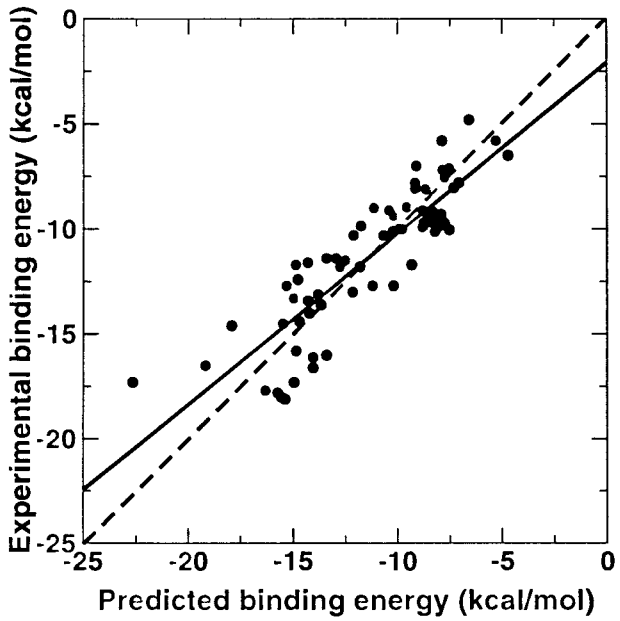

Fig. 4. The theoretically predicted binding free energy versus experimentally measured ones. The solid line is from linear regression fit, with a correlation coefficient of 0.87, a slope of 0.82. The theoretical results were shifted by a constant value of −4.7 kcal/mol, so that the intercept is zero. Dashed line indicates the location if there were a perfect agreement.

artificial interfaces. If one uses −9 kcal/mol as the cutoff value to distinguish the true from the fake interfaces, 93% of the dimers or monomers are assigned correctly. (The cutoff was chosen for maximizing the success rate. A similar method was used in other work to determine the cutoff value.[11,26]) This is comparable to the success rate of 95% by the LLS potential derived from the dimeric interfaces and 93% by a method of atomic contact vectors[26] but is superior to the rate of 86% by a sequence-based method[27] and the rate of 85–88% by a solvent-accessible surface area and a pair scoring function.[17] The success rate of the LLS potential derived from monomer structures, however, is significantly poorer. A success rate of 59% for a residue-level potential extracted from monomer structures was reported, compared to 86% for the same potential but trained by the interfacial regions of dimers.[11]

Binding Free Energy

The most critical test of the monomer DFIRE potential is its ability to predict binding free energy. A satisfactory prediction would suggest that the monomer DFIRE potential captures the essential physics behind protein–protein binding. The results of theoretically predicted binding free energies are compared with experimentally measured ones in Figure 4 and Table III. The correlation coefficient between the two sets of data is 0.867. More significantly, without introducing an additional parameter, the regression slope (0.82) is close to 1. [The slope becomes 0.87 if the most significant outlier (the point with the lowest predicted binding free energy) is removed. There is also a cluster of the binding free energies about 8 kcal/mol. These complexes contain a common receptor called oligopeptide binding protein (OPPA). Once we randomly select one of them and remove the rest from the data set, the slope (0.93) is further closer to 1.] There is an intercept of −4.7 kcal/mol. This systematic constant shift of theoretical results from experimental data is mostly caused by different standard states and reference volumes used in the theory and experiments.[28,29] For example, a conversion from experimental concentration unit of liter/molecule to the theoretical unit of Angstroms cubed/molecule will lead to a constant shift of −4.42 kcal/mol at T = 300 K. This is because the absolute binding free energy refers to the binding of B from a reference concentration of $1/V_{ref}$ ($-RT \ln K_{AB}/V_{ref}$),[28] where the association constant K_{AB} typically has a unit of volume in liter/molecule (i.e., V_{ref} =1 L/mol = 1660 Å^3/mol). Other factors such as distance cutoff and possible systematic errors in the potential also contribute to the constant shift. The RMSD ($\sqrt{\Sigma(G^{exp}_{bind} - G^{pred}_{bind})^2/N}$) between experimental data and theoretical prediction (after a −4.7 kcal/mol constant shift) is only 1.76 kcal/mol. Thus, the DFIRE-based monomer potential provides a reasonably accurate description of energetic and entropic contributions to the binding stability. This is remarkable considering the fact that many assumptions have been made in calculating the binding free energies. They include the rigid-body approximation and the lack of explicit treatment of long-range electrostatics and water molecules. The level of prediction accuracy reported here is comparable to or better than other small-scale studies (9–28 complexes) using distance-dependent knowledge-based potentials derived from interfacial structures[18,30] and other methods.[31–34] This is the first attempt, however, to evaluate the binding free energy using a monomer potential without additional adjustable parameters (except the constant shift of −4.7 kcal/mol).

DISCUSSION

Dependence on Database

The performance of DFIRE potential is mostly independent of the structural database used for extracting the

TABLE III. Comparison Between Experimentally Measured Binding Free Energies and Theoretically Predicted Ones

PDB ID[a]	Interface[b]	Exp.[c]	Theory[d]	PDB ID[a]	Interface[b]	Exp.[c]	Theory[d]
1hbs	ABCD/EFGH	−4.8[38]	−6.57	2tpi	ZI/S	−5.8[31]	−5.27
1tce	A/B	−5.8[39]	−7.86	1ak4	A/C	−6.5[40]	−4.70
1lck	A/B	−7.0[39]	−9.09	1b4z	A/B	−7.1[41]	−7.52
1b46	A/B	−7.2[41]	−7.82	2olb	A/B	−7.6[41]	−7.72
1lcj	A/B	−7.8[39]	−9.13	3tpi	Z/S	−7.8[42]	−7.07
1b3l	A/B	−8.0[41]	−7.30	1qka	A/B	−8.1[41]	−9.12
1b9j	A/B	−8.1[41]	−8.63	2pld	A/B	−9.0[39]	−11.11
1b58	A/B	−9.0[41]	−9.53	1sps	A/D	−9.1[39]	−8.77
1dkz	A/B	−9.1[43]	−10.38	1b3g	A/B	−9.2[41]	−8.33
1jeu	A/B	−9.3[41]	−7.92	1b3f	A/B	−9.4[41]	−8.54
1jev	A/B	−9.4[41]	−10.16	1ola	A/B	−9.5[44]	−8.45
1b5i	A/B	−9.6[41]	−8.06	1b05	A/B	−9.7[41]	−8.73
1b32	A/B	−9.7[41]	−8.40	1b52	A/B	−9.7[41]	−7.75
2er6	E/I	−9.8[44]	−11.75	1jet	A/B	−9.8[41]	−7.92
1b40	A/B	−9.9[41]	−8.77	1b51	A/B	−10.0[41]	−7.51
1qkb	A/B	−10.0[41]	−8.14	1nmb	HL/N	−10.0[45]	−9.90
2pcc	A/B	−10.0[46]	−9.75	1b5j	A/B	−10.1[41]	−8.19
1gua	A/B	−10.1[47]	−10.23	1dkg	AB/D	−10.3[48]	−10.67
1ycs	A/B	−10.3[49]	−12.11	1fdl	HL/Y	−11.4[50]	−12.95
1vfb	AB/C	−11.4[51]	−13.40	2jel	HL/P	−11.5[52]	−12.51
1abi	HL/I	−11.6[53]	−14.27	1ebp	A/C	−11.7[54]	−9.31
4sgb	E/I	−11.7[55]	−14.85	1jhl	HL/A	−11.8[56]	−12.75
1nsn	HL/S	−11.8[57]	−11.81	2kai	AB/I	−12.4[58]	−14.76
2sic	E/I	−12.7[59]	−15.29	3sgb	E/I	−12.7[60]	−10.18
1igc	HL/A	−12.7[61]	−11.19	1hwg	A/C	−13.0[62]	−12.15
1cse	E/I	−13.1[63]	−13.8	3hfm	HL/Y	−13.3[64]	−14.95
1ppf	E/I	−13.4[65]	−14.28	3hhr	A/C	−13.6[66]	−13.67
1tec	E/I	−14.0[63]	−14.21	3hfl	HL/Y	−14.5[31]	−15.45
1cho	E/I	−14.6[67]	−14.69	4htc	HL/I	−14.6[65]	−17.91
2sni	E/I	−15.8[68]	−14.84	3ssi	Symmetry[e]	−16.0[69]	−13.41
1acb	E/I	−16.1[70]	−14.04	1bth	HL/P	−16.5[62]	−19.16
1efn	A/B	−16.6[71]	−14.05	1brs	B/E	−17.3[72]	−14.95
1tbq	JK/S	−17.3[73]	−22.65	4tpi	Z/I	−17.7[74]	−16.29
1tpa	E/I	−17.8[31]	−15.72	1dfj	E/I	−18.0[75]	−15.55
2ptc	E/I	−18.1[76]	−15.37				

[a]The database does not include proteins with metal atoms and other non-amino-acid components at the interface. Some of the data were originately collected by Brooijmans et al.[77]
[b]The chain IDs that make the interface.
[c]Experimental results (in kcal/mol).
[d]Predicted values (in kcal/mol) by the DFIRE monomer potential, shifted by a constant value of −4.7 kcal/mol.
[e]The second component of complex 3ssi was generated with the symmetry axis provided by PDB file.

potential. The database of dimeric structures[11] is also used to generate the DFIRE dimer potential. This database contains 768 protein dimer structures (617 homodimers and 151 heterodimers), with no more than 35% sequence identity and a resolution better than 3 Å. The entire complex structure including interface is used in generating the statistics on pair distributions. This differs from the usage of the original 1011 structural database for statistics, where only the structure of one chain from the Protein Data Bank (PDB) file was used. All the results are essentially the same. The DFIRE dimer (monomer) potential has a success rate of 95% (100%) in the 21 Sternberg and Vakser docking decoys (based on the first rank), 29/31 (29/31) in the Rosetta docking decoys [based on the Z score(native)], 94% (93%) in interface discrimination, and a correlation coefficient of 0.85 (0.87) between theoretical results and experimental data of binding free energies. We also obtained the DFIRE potential based on the unphysical restriction of the database to interfacial structures. The performance of this potential is significantly worse than that of the DFIRE-based monomer potential.

Dependence on the Distance Cutoff for Interface

The performance depends somewhat on the distance cutoff used to define interacting interfacial residues. We find that a cutoff of 4.5 Å used by Lu et al.[11] is also the optimal value for the performance of the DFIRE-monomer potential. For example, the success rate for structure selections in the 21 decoy sets is reduced from 21/21 with the cutoff to 17/21 without cutoff. The correlation coefficient between theoretical results and experimental data of binding free energies is reduced from 0.87 to 0.82. The reduction of performance for a larger cutoff suggests that DFIRE potential has overestimated the importance of "noninterfacially interacting" residues. This points out that the DFIRE potential at the intermediate range ($r <$

r_{cut}) is inaccurate. This inaccuracy may result from the uniform cutoff of the DFIRE potential at $r = 14.5$ Å. The cutoff distance of 14.5 Å was chosen because the distance dependence for the number of pairs of ideal-gas points in finite spheres starts to systematically deviate from $r^{1.61}$ for $r > 15$ Å.[21]

Prospect for Docking Structure Prediction

Although the structure-derived DFIRE potential makes a significant improvement over other structure-derived potential, there is still room for further improvement. For example, the success rate based on the first rank for the native structure is only 29% (9/31) for the DFIRE-based monomer potential in Rosetta docking decoys. (For KMB,[10] the ranking information was not reported.) Nevertheless, most native structures are ranked among the tops 400 decoys [61% (19/31) and 71% (22/31) are within top 5 and top 10, respectively]. Thus, the Rosetta docking decoy sets are more challenging than the 21 dimer/trimer decoy sets. One immediate improvement to the current application of DFIRE potential to docking is to allow some flexibility in side-chain positions. The work in this area is currently in progress.

CONCLUSIONS

Intra- and interprotein interactions involve the same physical, water-mediated potential of mean force between amino acid residues. This physical property, masked under different amino acid preferences within a single protein and between protein–protein complexes, is difficult to realize in commonly used knowledge-based potentials using statistically average reference states. This work suggests that a simple physical reference state of ideal gas not only makes the structure-derived potential more accurate but also makes it more physical, so that a unified potential of mean force can be built for both folding and binding studies. The potential offers an alternative to many physical-based energy functions in which the contribution of entropy is difficult to calculate. The DFIRE-based potential is expected to be useful in structure selections,[21] stability prediction,[21] loop prediction[21], fold recognition via threading,[35,36] protein (peptide)-protein docking,[37] and the prediction of binding free energy with improved accuracy.

ACKNOWLEDGMENTS

We gratefully thank Professor David Baker and Dr. Alex Morozov for providing us the Rosetta docking decoy sets, to Dr. Hannes Ponstingl for the list of hypothetical dimers' structure, and to Professor Luhua Lai for binding free-energy data of several protein complexes.

REFERENCES

1. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
2. Kaminski GA, Stern HA, Berne BJ, Friesner R. Development of a polarizable force field for proteins via *ab initio* quantum chemistry: first generation model and gas phase tests. J Comp Chem 2002;23:1515–1531.
3. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
4. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
5. 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.
6. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
7. Chong LT, Dempster SE, Hendsch ZS, Lee L-P, Tidor B. Computation of electrostatic complements to proteins: a case of charge stabilized binding. Protein Sci 1998;7:206–210.
8. Wang W, Kollman PA. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. J Mol Biol 2000;303:567–582.
9. Sheinerman F, Norel R, Honig B. Electrostatics aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
10. Kortemme T, Morozov A, Baker D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. J Mol Biol 2003;326:1239–1259.
11. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein–protein interactions. Biophys J 2003; 84:1895–1901.
12. Glaser F, Sternberg D, Vakser I, Ben-Tal N. Residue frequencies and pairing preferences at protein–protein interfaces. Proteins 2001;43:89–102.
13. Ofran Y, Rost B. Analyzing six types of protein–protein complexes. J Mol Biol 2003;325:377–387.
14. Moont G, Gabb H, Sternberg M. Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 1999;35:364–373.
15. Robert C, Janin J. A soft, mean field potential derived from crystal contacts for predicting protein–protein interactions. J Mol Biol 1998;283:1037–1047.
16. Mitchell JBO, Laskowski RA, Alex A, Thornton JM. BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential. J Comp Chem 1999;20:1165–1176.
17. Ponstingl H, Henrick K, Thornton J. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
18. Jiang L, Gao Y, Mao F, Liu Z, Lai L. Potential of mean force for protein–protein interaction studies. Proteins 2002;46:190–196.
19. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998;275:895–916.
20. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 2001; 44:223–232.
21. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714–2726; 2003;12:2121.
22. Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. Proteins 2004;54:315–322.
23. Lu L, Lu H, Skolnick J. Multiprospector: an algorithm for the prediction of protein–protein interactions by multimeric threading. Proteins 2002;49:350–364.
24. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
25. Camacho CJ, Gatchell DW, Kimura SR, Vajda S. Scoring docked conformations generated by rigid-body protein–protein docking. Proteins 2000;40:525–537.
26. Mintseris J, Weng Z. Atomic contact vectors in protein–protein recognition. Proteins 2003;53:629–639.
27. Elock A, McCammon J. Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci USA 2001;98:2990–2994.
28. Gilson MK, Given JA, Bush BL, McCammon JA. The statistical–thermodynamic basis for computation of binding affinities: a critical review. Biophys J 1997;72:1047–1069.
29. Luo H, Sharp K. On the calculation of absolute macromolecular binding free energies. Proc Natl Acad Sci USA 2003;99:10399–10404.
30. Mitchell JBO, Laskowski RA, Alex A, Foster MJ, Thornton JM. BLEEP potential of mean force describing protein-ligand interactions: II. Calculation of binding energies and comparison with experimental data. J Comp Chem 1999;20:1177–1185.

31. Horton N, Lewis M. Calculation of the free energy of association for protein complexes. Protein Sci 1992;1:169–181.
32. Vajda S, Weng Z, Rosenfeld R, Delisi C. Effect of conformational flexibility and solvation on receptor-ligand binding free energies. Biochemistry 1994;33:13977–13988.
33. Xu D, Lin SL, Nussinov R. Protein binding versus protein folding: the role of hydrophilic bridges in protein associations. J Mol Biol 1997;265:68–84.
34. Zhou H, Zhou Y. The stability scale and atomic solvation parameters extracted from 1023 mutation experiments. Proteins 2002;49: 483–492.
35. Meller J, Elber R. Protein recognition by sequence-to-structure fitness: bridging efficiency and capacity of threading models. Adv Chem Phys 2002;120:77–130.
36. Zhou H, Zhou Y. Single body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004. Forthcoming.
37. Smith GR, Sternberg MJE. Prediction of protein–protein interactions by docking methods. Curr Opin Struct Biol 2002;12:28–35.
38. Ross PD, Hofrichter J, Eaton WA. Thermodynamics of gelation of sickle cell deoxyhemoglobin. J Mol Biol 1977;115:111–134.
39. Zhou Y, Abagyan AR. How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains. Fold Des 1998;3:513–522.
40. Gamble TR, Vajdos FF, Yoo S, Worthylake DK, Houseweart M, Sundquist WI, Hill CP. Crystal structure of human cyclophilin a bound to the amino-terminal domain of HIV-1 capsid. Cell 1996;87: 1285–1294.
41. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of oppA-peptide complexes to relate structure to binding thermodynamics. J Med Chem 2002;45:4828–4837.
42. Vincent JP, Lazdunski M. Pre-existence of the active site in zymogens, the interaction of trypsinogen with the basic pancreatic trypsin inhibitor (Kunitz). FEBS Lett 1976;63:240–244.
43. Krystek S, Stouch T, Novonty J. Affinity and specificity of serine endopeptidase- protein inhibitor interactions: empirical free energy calculations based on X-ray crystallographic structures. J Mol Biol 1993;234:661–679.
44. Puvanendrampillai D, Mitchell J. Protein Ligand Database (PLD): Additional understanding of the nature and specificity of protein–ligand complexes. Bioinformatics 2003;19:1856–1857.
45. 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.
46. Corin AF, McLendon G, Zhang QP, Hake RA, Falvo J, Lu KS, Ciccarelli RB, Holzschu D. Effects of surface amino acid replacements in cytochrome c peroxidase on complex formation with cytochrome c. Biochemistry 1991;30:11585–11595.
47. Nassar N, Horn G, Herrmann C, Block C, Janknecht R, Wittinghofer A. Ras/Rap effector specificity determined by charge reversal. Nat Struct Biol 1996;3:723–729.
48. Harrison CJ, HayerHartl M, DiLiberto M, Hartl FU, Kuriyan J. Crystal structure of the nucleotide exchange factor GrpEbound to the ATPase domain of the molecular chaperone DnaK. Science 1997;276:431–435.
49. Gorina S, Pavletich NP. Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. Science 1996;274:1001–1005.
50. Arnon R. Synthetic peptides as the basis for vaccine design. Mol Immunol 1991;28:209–215.
51. Verhoeyen M, Milstein C, Winter G. Reshaping human antibodies: grafting an antilysozyme activity. Science 1988;239:1534–1536.
52. Smallshaw JE, Brokx S, Lee JS, Waygood EB. Determination of the binding constants for three HPr-specific monoclonal antibodies and their fab fragments. J Mol Biol 1998;325:765–774.
53. Qiu X, Padmanabhan KP, Carperos VE, Tulinsky A, Kline T, Maraganore JM, Fenton JW II. Structure of the hirulog 3-thrombin complex and nature of the S′ subsites of substrates and inhibitors. Biochemistry 1992;31:11689–11697.
54. Lee CH, Saksela K, Mirza UA, Chait BT, Kuriyan J. Crystal structure of the conserved core of HIV-1 Nef complexed with a src family SH3 domain. Cell 1996;85:931–942.
55. Hass GM, Ryan CA. Carboxypeptidase inhibitor from potatoes. Methods Enzymol 1981;80:779–790.
56. Chitarra V, Alzari PM, Bentley GA, Bhat TN, Eisele J, Houdusse A, Lescar J, Souchon H, Poljak RJ. Three-dimensional structure of a heteroclitic antigen–antibody cross-reaction complex. Proc Natl Acad Sci USA 1993;90:7711–7715.
57. Smith AM, Benjamin DC. The antigenic surface of staphylococcal nuclease: 2. Analysis of the n-1 epitope by site-directed mutagenesis. J Immunol 1991;146:1259–1264.
58. Dietl T, Huber C, Geiger R, Iwanaga S, Fritz H. Inhibition of porcine glandular kallikreins by structurally homologous proteinase inhibitor of the Kunitz type. Hoppe-Seyler's Z Physiol Chem 1979;360:67–71.
59. Uehara Y, Tonomura B, Hiromi K. Direct fluorometric determination of a dissociation constant as low as 10^{-10}m for the subtilisin BPN′-protein proteinase inhibitor (*Streptomyces subtilisin* inhibitor) complex by a single photon counting technique. J Biochem 1978;84:1195–1202.
60. Bigler TL, Lu W, Park SJ, Tashiro M, Wieczorek M, Wynn R, Laskowski MJ. Binding of amino acid side chains to preformed cavities: interaction of serine proteinases with turkey ovomucoid third domains with coded and noncoded P1 residues. Protein Sci 1993;2:786–799.
61. Sjobring U, Bjorck L, Kastern W. Streptococcal protein G gene structure and protein bind properties. J Biol Chem 1991;266:399–405.
62. Guinto ER, Ye J, Bonniec BFL, Esmon CT. Glu192 → Gln substitution in thrombin yields an enzyme that is effectively inhibited by bovine pancreatic trypsin inhibitor and tissue factor pathway inhibitor. J Biol Chem 1994;269:18395–18400.
63. Seemuller U, Fritz H, Euliz M. Eglin elastase-cathepsing inhibitor from leeches. Methods Enzymol 1981;80:804–816.
64. Padlan E, Silverton E, Sheriff S, Cohen G, Smith-Gill S, Davies D. Structure of an antibody–antigen complex: crystal structure of the hyhel-10 fab-lysozyme complex. Proc Nat Acad Sci USA 1989;86: 5938–5942.
65. Bode W, Mayr I, Baumann U, Huber R, Stone S, Hofsteenge J. The refined 1.9 angstroms crystal structure of human alpha-thrombin: interaction with D-PHE-PRO-ARG chloromethylketone and significance of the TYR-PRO-PRO-TRP insertion segment. EMBO J 1989;8:3467–3475.
66. Vos AD, Ultsch M, Kossiakoff A. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. Science 1992;255:306–312.
67. Lu WY, Qasim MA, Laskowski M, Kent SBH. Probing intermolecular main chain hydrogen bonding in serine proteinase-protein inhibitor complexes: chemical synthesis of backbone-engineered turkey ovomucoid third domain. Biochemistry 1997;36:673–679.
68. Svendsen I, Jonassen I, Hejgaard J, Boisen S. Amino acid sequence homology between a serine protease inhibitor from barley *Hordeum vulgare* cultivar Hiproly and potato inhibitor I. Carsberg Res Commun 1980;45:389–395.
69. Akasaka K, Fujii S, Hayashi F, Rokushika S, Hatano H. A novel technique for the detection of dissociation–association equilibrium in highly associable macromolecular systems. Biochem Int 1982;5: 637–642.
70. Qasim MA, Ganz PJ, Saunders CW, Bateman KS, James MNG, Laskowski M. Interscaffolding additivity: association of P-1 variants of eglin C and of turkey ovomucoid third domain with serine proteinases. Biochemistry 1997;36:1598–1607.
71. Clackson T, Ultsch MH, Wells JA, deVos AM. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. J Mol Biol 1998;277:1111–1128.
72. Hartley RW. Directed mutagenesis and barnase-barstar recognition. Biochemistry 1993;32:5978–5984.
73. van deLocht A, Lamba D, Bauer M, Huber R, Friedrich T, Kroger B, Hoffken W, Bode W. Two heads are better than one: crystal structure of the insect derived double domain Kazal inhibitor rhodniin in complex with thrombin. EMBO J 1995;14:5149–5157.
74. 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.
75. Kobe B, Deisenhofer J. A structural basis of the inacteractions between leucine-rich repeats and protein ligands. Nature 1995;374: 183–186.
76. Vincent JP, Lazdunski M. Trypsin-pancreatic inhibitor association: dynamic of the interaction and role of disulfide bridges. Biochemistry 1972;11:2967–2977.
77. Brooijmans N, Sharp KA, Kuntz ID. Stability of macromolecular complexes. Proteins 2002;48:645–653.