A Knowledge-Based Energy Function for Protein−Ligand, Protein−Protein, and Protein−DNA Complexes

Chi Zhang,[†,‡] Song Liu,[†,‡] Qianqian Zhu,[†] and Yaoqi Zhou*[,†,§]

Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, New York 14214, and Department of Macromolecular Science, The Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai 200433, China

Received August 19, 2004

We developed a knowledge-based statistical energy function for protein−ligand, protein−protein, and protein−DNA complexes by using 19 atom types and a *d*istance-scale *f*inite *i*deal-gas *re*ference (DFIRE) state. The correlation coefficients between experimentally measured protein−ligand binding affinities and those predicted by the DFIRE energy function are around 0.63 for one training set and two testing sets. The energy function also makes highly accurate predictions of binding affinities of protein−protein and protein−DNA complexes. Correlation coefficients between theoretical and experimental results are 0.73 for 82 protein−protein (peptide) complexes and 0.83 for 45 protein−DNA complexes, despite the fact that the structures of protein−protein (peptide) and protein−DNA complexes were not used in training the energy function. The results of the DFIRE energy function on protein−ligand complexes are compared to the published results of 12 other scoring functions generated from either physical-based, knowledge-based, or empirical methods. They include AutoDock, X-Score, DrugScore, four scoring functions in Cerius 2 (LigScore, PLP, PMF, and LUDI), four scoring functions in SYBYL (F-Score, G-Score, D-Score, and ChemScore), and BLEEP. While the DFIRE energy function is only moderately successful in ranking native or near native conformations, it yields the strongest correlation between theoretical and experimental binding affinities of the testing sets and between rmsd values and energy scores of docking decoys in a benchmark of 100 protein−ligand complexes. The parameters and the program of the all-atom DFIRE energy function are freely available for academic users at http://theory.med.buffalo.edu.

Introduction

The common bottleneck behind computational studies on protein-structure prediction,[1] structure-based ligand design,[2,3] protein−protein docking prediction,[4] and prediction of protein−DNA interaction[5−7] is the lack of a precise and reliable free energy function that describes the water-mediated atomic interaction between amino acid residues and other molecules. Obtaining an accurate energy function is challenging because the stability of proteins themselves as well as specific binding between proteins and other molecules is the result of the delicate balance among several different types of interactions,[8,9] such as van der Waals' (packing), hydrophobic (solvent-induced), hydrogen-bonding, polar, and charge−charge interactions. Additional complications arise from the fact that complementarity in shape, hydrogen-bonding, polar, and charge−charge interactions is required for structural specificity of folding and binding (molecular recognition). Many approaches have been used to obtain approximate energy functions. The existing energy functions for protein−ligand (and/or protein−protein) interactions can be classified as physical-based force fields,[10−13] empirical scoring functions,[14−18] and knowledge-based statistical potentials.[19−27]

In this paper, we focus on knowledge-based statistical potentials, which derive energy functions from statistical analysis of known structures of proteins or protein−ligand (or protein−protein) complexes. Knowledge-based potentials are attractive because they are simple to construct and easy to use. However, existing statistical energy functions are often associated with some properties that are not physical. For example, long-range repulsion is observed[28] between hydrophobic residues in the statistical energy function based on the commonly used Sippl approximation.[29] In addition, statistical potentials are strongly database dependent. For example, the potential extracted from all-α protein structures is quantitatively different from that extracted from all-β protein structures.[30] The structural database of single-chain proteins and the structural database of dimeric interfaces also yielded different statistical potentials for folding and binding.[26,31] To our knowledge, the database dependence of statistical potentials for protein−ligand interactions is not yet tested. The database dependence reflects the fact that proteins are inhomogeneous mixture of amino acid residues and the different compositions of amino acid residues in different structural databases lead to different statistical outcome (i.e. the energy function). This strong database dependence is often used to produce a system-specific statistical energy function to improve the performance of the energy function for a given system. For example, the knowledge-based potential based on the structural

* Corresponding author. Phone: (716) 829-2985. Fax: (716) 829-2344. E-mail: yqzhou@buffalo.edu.

[†] State University of New York at Buffalo.

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

[§] Fudan University.

10.1021/jm049314d CCC: $30.25 © 2005 American Chemical Society
Published on Web 02/16/2005

database of dimeric interfaces improves the success rate of selecting native complex structures from decoys over the potential based on the structural database of monomeric proteins.[26] The database dependence, however, might have limited the accuracy of statistical energy functions because the dependence does not occur in a real physical interaction (assuming that pairwise interactions are dominant). After all, the same underlying physical interaction (the water-mediated interaction between amino acid residues) is responsible for folding and binding and for the formation of α-helices and the formation of *β*-sheets.

Various knowledge-based potentials differ in choice of reference state,[32] which is used for estimating statistics in the absence of any interactions. A reference state is required for obtaining the net contribution of atomic interactions to the statistical results based on known structures by removing the contribution from a zero-interaction reference state. Most existing reference states are represented by a state that was averaged either over different atom types[19-21,25,26,29] or over distance.[22,23,32] Recently, we introduced a new knowledge-based potential based on a distance-scaled, finite, ideal-gas reference (DFIRE) state.[33] Remarkably, this reference state yields a potential of mean-force that no longer possesses some unphysical characteristics associated with other statistical potentials. It was shown that the accuracy of DFIRE-based potential is insensitive to whether residues at the surface or inside core of proteins are treated as separate residue types.[33] More importantly, the new structure-derived potential can quantitatively reproduce the likelihood of a residue to be buried (i.e. the composition difference of amino acid residues between core and surface).[34] The potential also produces a stability scale of amino acid residues in quantitative agreement with that independently extracted from mutation experimental data.[34] Moreover, the "monomer" potential (derived from single-chain proteins) is found to be successful in discriminating native structure against docking decoys, distinguishing true dimeric interface from crystal interfaces, and predicting binding free energy of protein–protein and protein–peptide complexes.[35] In addition, the DFIRE potential is less dependent on the structural database used for training than two other commonly used statistical potentials.[36] The independence of the performance on amino acid composition suggests that the DFIRE-based potential captures the essence of the common physical interaction masked under different compositions of amino acid residues on the surface, core and interface of proteins.

The success of the DFIRE-based statistical energy function for predicting protein–protein (peptide) binding affinity[35] provides the incentive for extending the DFIRE-based energy function for predicting binding affinity between protein and organic molecule. The original DFIRE energy function was derived on the basis of residue-specific atom types. That is, each atom in different amino acid residues was treated as an atom type. This leads to a total of 167 atom types for amino acid residues alone. Here, we use only 19 atom types to characterize the interaction not only between protein and protein but also between protein and ligand and between protein and DNA. The resulting energy function is compared to 12 different scoring functions for protein–ligand interactions. The energy function is also applied to predict protein–protein and protein–DNA binding affinities.

Table 1. List of 19 Atom Types Used in Protein–Ligand, Protein–Protein, and Protein–DNA Interactions

atom type	atoms	atom type	atoms
C.2	sp^2 carbon	C.3	sp^3 carbon
C.ar	aromatic carbon	C.cat	other carbon
N.2	sp^1,sp^2,sp^3 nitrogen	N.4	quaternary nitrogen
N.am	amide nitrogen	N.ar	aromatic nitrogen
N.p13	trigonal nitrogen	O.2	sp^2 oxygen
O.3	sp^3 oxygen	O.co2	carboxy oxygen
P.3	all phosphorus atoms	S.3	all sulfur atoms except sulfone sulfur
S.o2	sulfone sulfur	F	fluorine
Cl	chlorine	Br	bromine
		Met	all metal atoms and other atoms not listed above

Methods

DFIRE-Based Potential. The derivation of equations and the method for extracting the DFIRE-based potential using a structure database have been described previously.[33] 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[33]

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

where R is the gas constant, $T = 300$ K, $N_{\mathrm{obs}}(i,j,r)$ is the number of (i,j) pairs within the distance shell $r(r - \Delta r/2$ to $r + \Delta r/2)$ observed in a given structure database, $r_{\mathrm{cut}} = 14.5$ Å, and $\Delta r(\Delta r_{\mathrm{cut}})$ is the bin width at $r(r_{\mathrm{cut}})$. ($\Delta r = 2$ 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 exponent α is found to be 1.61 on the basis of a state of uniformly distributed points in finite spheres.[33]

Atom Types. Unlike the residue-specific atom types used in our previous work, we use 19 atom types to cover protein–ligand, protein–protein, and protein–DNA interactions. The atom types and definitions are shown in Table 1. These atom types are derived from those used in the program SYBYL.[37] All metal atoms are unified as one atom type: Met. Clearly, this small number of atom types is a crude approximation. We defer finer classification to future work.

Training Structural Databases. To calculate $N_{\mathrm{obs}}(i,j,r)$, one needs a structural database for training. The DFIRE potential was trained from a set of 200 protein–ligand structures, collected by Wang et al.[18] to train their X-Score scoring function. This set is a dataset of high resolution (<3.0 Å) of structures of proteins complexed with small organic noncovalently binding ligands ($M_{\mathrm{W}} < 1000$) but without additional binding cofactor. The dataset contains 70 different types of proteins, whose protein–ligand binding affinities vary over 10 orders of magnitude.

To test the dependence of the DFIRE potential on the training structural database, we also used one additional structural database of 1011 single-chain, nonhomologous (less than 30% homology) proteins with resolution <2 Å (http://chaos.fccc.edu/research/labs/dunbrack/culledpdb.html).[38] This database was used to generate the residue-specific all-atom DFIRE-based energy function.[33] It is employed here to test the effect of the number of atom types on the accuracy of the DFIRE-based energy function. All results reported here are based on the 200 protein–ligand data set, unless otherwise stated.

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 all atomic pairs. The binding free energy of a dimer AB is obtained as follows:

$$\Delta G_{\mathrm{bind}} = G_{\mathrm{complex}} - (G_{\mathrm{A}} + G_{\mathrm{B}}) \qquad (3)$$

Since the structures of monomers are approximated as rigid bodies and the atoms at the binding interface contribute most to ΔG_{bind}, eq 3 can be further simplified to

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

where the summation is over any two atoms belonging to an "interacting" atomic pair from different chains at the interface. An interacting atomic pair is one pair of heavy atoms within 9.5 Å of each other. Other cutoff values were also tested. See discussion for details.

Testing Sets for Binding Affinity and Docking. We compare the performance of 11 scoring functions with that of the DFIRE energy function in predicting protein−ligand binding affinity and docking structure selections based on the benchmarks established by Wang et al.[39] The 11 scoring functions are Autodock,[11] LigScore, PLP,[15,40] PMF,[22] LUDI,[14] F-Score,[16] G-Score,[10] D-Score,[13] ChemScore,[17] DrugScore,[23] and X-Score.[18,39] Among these scoring functions, PMF and DrugScore are statistical potentials as DFIRE.

In addition, we compared the DFIRE energy function with another statistical potential, BLEEP,[20,21] using the data downloaded from the Protein Ligand Database (PLD) (http://www-mitchell.ch.cam.ac.uk/pld/).

To further test the DFIRE energy function, experimental binding free energies of 82 protein−protein (peptide) complexes and 45 protein−DNA complexes all with known three-dimensional structures were collected from the literature.

Results

The Potential. The DFIRE-based energy functions between atom types O.3 and S.3 and between C.2 and N.4 are shown in Figure 1 as an example. Two sets of curves are the two DFIRE potentials obtained from the training structural database of 1011 proteins and that of 200 protein−ligand complexes, respectively. There are some differences between the two potentials (near the hard-repulsive core region, in particular). However, as we shall see below, the differences only have a minor effect on the overall accuracy of the potential for predicting binding affinity.

Prediction of Protein−Ligand Binding Affinity Based on X-Score Training and Testing Sets. The abilities of the DFIRE and X-Score energy functions[18] to predict binding affinity are compared in Figure 2. It should be noted that the DFIRE energy and X-Score energy functions shared the same training and testing sets. However, different information was used. The DFIRE energy used the structural information of the 200 protein−ligand complexes, whereas the X-Score used binding affinities of the 200 complexes directly for least-squares multivariate regression. In Figure 2, the experimentally measured binding affinities of 30 testing protein−ligand complexes are compared to those predicted by the DFIRE energy function and by X-Score. The correlation coefficients given by the DFIRE energy function are 0.62 for the training set and 0.63 for the testing set, respectively. The corresponding numbers given by X-Score are 0.76 and 0.56, respectively.[18] Thus, the performance of the X-Score is more dependent on the data set while the performance of DFIRE is robust.

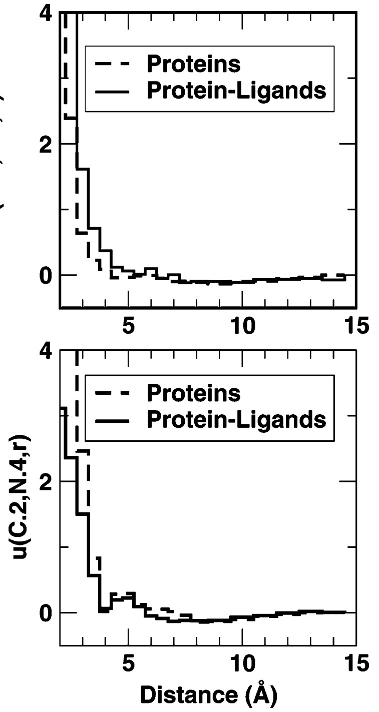

Figure 1. The DFIRE-based energy functions between O.3 and S.3 and between C.2 and N.4 as a function of distance. Solid and dashed lines are the energy functions trained with the structure database of protein−ligand complexes and that of single-chain proteins, respectively.

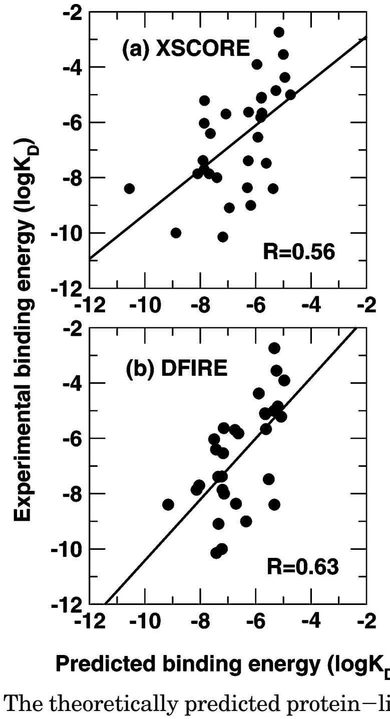

Figure 2. The theoretically predicted protein−ligand binding free energy versus experimentally measured ones for 30 complexes in the X-Score testing dataset.[18] To facilitate comparison, the DFIRE energy function was scaled with a constant scaling factor of 0.0051 and shifted by −3.84 in log K_{D} units based on the training set. The solid line is from the regression fit. (a) XSCORE and (b) DFIRE.

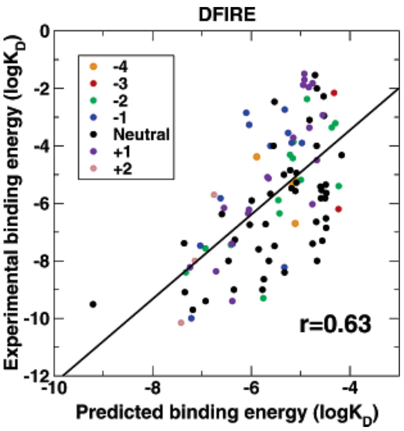

Figure 3. The theoretically predicted protein−ligand binding free energy versus experimentally measured ones for the 100 high-resolution protein−ligand set. The DFIRE energy function was scaled and shifted as in Figure 2. Ligands with different charges are shown in different colors as labeled.

Table 2. Correlation Coefficients between Theoretically Predicted and Experimentally Measured Binding Affinities for 100 Protein−Ligand Complexes

scoring functions	corr coeff	scoring functions	corr coeff
DFIRE[a]	0.63[b] (0.66[c])	SYBYL/ChemScore[d]	0.47
X-Score[d]	0.64[b]	Cerius2/LUDI[d]	0.37
Cerius2/PLP[d]	0.55	Cerius2/PMF[d]	0.40
DrugScore[d]	0.57	Cerius2/LigScore[d]	0.35
SYBYL/G-Score[d]	0.56	SYBYL/F-Score[d]	0.30
SYBYL/D-Score[d]	0.48	AutoDock[d]	0.05

[a] This work. [b] Many of the 100 complexes were used in training X-Score and DFIRE. [c] Results from DFIRE training by the structures of 1011 single-chain proteins. [d] Results from Wang et al.[39] Here we reported their results in term of correlation coefficients.

Table 3. Success Rates for Selecting Native or Near-Native Conformations under Different rmsd Cutoffs Given by 12 Different Scoring Functions[a]

scoring function	≤1.0 Å	≤1.5 Å	≤2.0 Å	≤2.5 Å	≤3.0 Å
DFIRE	37	52	58	61	64
Cerius2/PLP	63	69	76	79	80
SYBYL/F-score	56	66	74	77	77
Cerius2/LigScore	64	68	74	75	76
DrugScore	63	68	72	74	74
Cerius2/LUDI	43	55	67	67	67
X-Score	37	54	66	72	74
AutoDock	34	52	62	68	72
Cerius2/PMF	40	46	52	54	57
SYBYL/G-score	24	32	42	49	56
SYBYL/ChemScore	12	26	35	37	40
SYBYL/D-Score	8	16	26	30	41

[a] All results except DFIRE are taken from Wang et al.[39]

Table 4. Number of Complexes Whose Spearman Correlation Coefficients between rmsd Values and Energy Scores of Docking Decoys Are Greater than a Cutoff Value

scoring functions	$R_s \geq 0.8$[a]	$R_s \geq 0.6$[a]	$R_s \geq 0.4$[a]
DFIRE[b]	23	63	88
X-Score[c]	19	53	77
DrugScore[c]	21	46	73
AutoDock[c]	12	42	71
Cerius2/PLP[c]	13	39	67
SYBYL/D-Score[c]	9	39	67
Cerius2/PMF[c]	21	38	61
Cerius2/LUDI[c]	8	37	66
SYBYL/F-Score[c]	9	34	72
SYBYL/G-Score[c]	6	28	56
Cerius2/LigScore[c]	4	26	49
SYBYL/ChemScore[c]	1	16	41

[a] The Spearman correlation coefficient between theoretically predicted binding free energies and rmsd values given by 12 different energy scoring functions as in ref 39. [b] This work. [c] From ref 39.

The performance of DFIRE is further tested by a set of 100 protein−ligand binding data collected by Wang et al.[39] This data set is a high-resolution subset of the training and testing data sets of 230 protein−ligand complexes. Results are shown in Figure 3. The correlation coefficient for this subset between experimentally measured and DFIRE-predicted binding affinities is 0.63. This correlation coefficient is only lower than that of X-Score but is higher than those of all other 10 scoring functions, as shown in Table 2. The correlation coefficients given by the latter 10 scoring functions range from 0.05 to 0.57. However, this set is not an independent testing set for the performance of either DFIRE or X-Score, because most complexes in the set of 100 complexes were used in training X-Score and DFIRE. To address this problem, we also applied the DFIRE energy function trained by 1011 single-chain proteins to this 100 protein−ligand complexes. The new DFIRE energy function achieved a similarly strong correlation between theoretical and experimental binding affinities (with a correlation coefficient of 0.66). This confirms that the performance of DFIRE is relatively independent of the structural database used for training. Moreover, it suggests that the DFIRE energy function gives the highest correlation coefficient among 11 scoring functions (except X-Score) for an independent testing set of 100 complexes.

Protein−Ligand Docking Decoys. It is important to have a statistically significant correlation between experimentally measured binding affinities and theoretical values predicted from known complex structures. However, in practice, the structures of complexes often are not known. Thus, it is important to apply an energy function directly to docking decoys.

Docking decoys for the above-mentioned 100 protein−ligand complexes are generated by Wang et al.[39] using the program AutoDock 3.0.[11] Each ligand has 101 docked conformations (including native conformations). Table 3 shows the success rates given by 12 different energy functions in selecting native or near-native conformations. The DFIRE energy function gives only a modest success rate similar to that of AutoDock.

Another method to characterize the ability to detect near native conformation is to calculate the correlation coefficient between the energy and rmsd values of the decoy conformations. Table 4 lists the number of protein−ligand complexes whose correlation coefficients are greater than or equal to a given number. (The Spearman correlation coefficient is used here for a convenient comparison with previous work.[39]) The results of 12 different energy functions are shown. The DFIRE energy function has a higher number of significant correlations than other energy functions. For example, the number of complexes with $R \geq 0.6$ is 63 for DFIRE but is 53 for the next best (X-Score).

One problem often encountered in structure-based drug design is how to predict protein binding affinity without knowing the structures of protein−ligand complexes. To test the DFIRE energy function on this aspect, we use the lowest energy of docking decoys in a

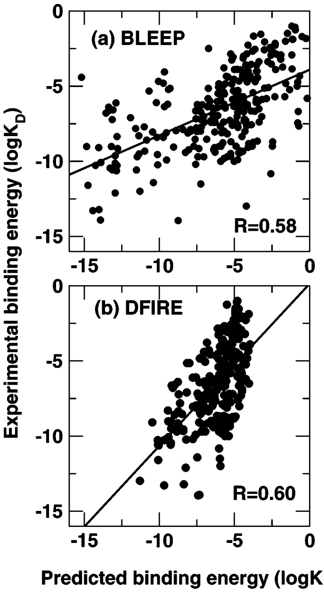

Figure 4. The theoretically predicted protein−ligand binding free energy versus experimentally measured ones for PLD dataset (268 complexes). The DFIRE energy function was scaled and shifted as in Figure 2. Solid lines are from the regression fit. (a) BLEEP and (b) DFIRE. The range of the predictions from DFIRE appears to be narrower than that of experimental data because the scaling factor and shift were from the training set.

protein−ligand complex as the predicted binding affinity, regardless if the complex structure is close or not to the native complex structure (all native structures are removed from the decoy sets). The correlation coefficient between experimental results and the lowest binding affinities of docking decoys is 0.64. Thus, the accuracy for predicting binding affinities does not change much even though most structures with the lowest DFIRE energy scores are not close to the native complex structures (Table 3).

Prediction of Protein−Ligand Binding Affinities Based on the PLD Set. We downloaded the entire data set from PLD[41] (the version of Jan. 10, 2004), which contains the PDB ID, experimental binding affinity, and the binding affinity predicted by the program BLEEP. Only protein−ligand complexes that have the experimental binding affinities were kept. It contains 268 complexes (a list is provided in http://theory.med.buffalo.edu). Figure 4 compares the experimentally measured binding affinities for the 268 complexes with those predicted by DFIRE and BLEEP. The correlation coefficient is 0.60 for DFIRE and 0.58 for BLEEP. Figure 4 further indicates that BLEEP's predictions deviate more from the regression line than DFIRE's predictions. The root-mean-squared deviation between theoretically predicted and experimentally measured binding affinities (in log K_D units) is 2.41 for DFIRE and 2.95 for BLEEP.

Among 268 complexes, 139 complexes are in the training set of the 200 protein−ligand complexes for the DFIRE energy function. The rest (129 complexes) can serve as an independent testing set for DFIRE. The correlation coefficient for this testing set is 0.64 by DFIRE and 0.59 by BLEEP. Thus, the overall accuracy of binding affinities predicted by DFIRE does not change for the training or testing set as demonstrated in the performance of DFIRE in X-Score training and testing benchmarks.

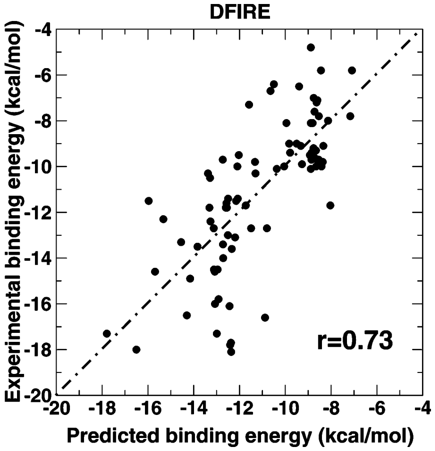

Figure 5. The theoretically predicted protein−protein (peptide) binding free energy versus experimentally measured ones. The line is from linear regression fit with a correlation coefficient of 0.73. To facilitate comparison, the DFIRE energy function was scaled with a constant scaling factor of 0.0059 and shifted by −5.35 in kcal/mol. The dashed line is from the regression fit. It is same as the $x = y$ line because of artificial scaling employed for clear view.

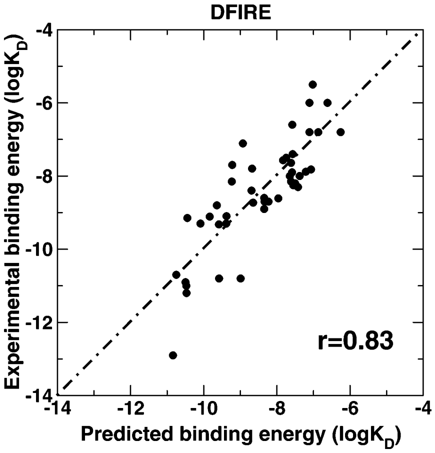

Figure 6. The theoretically predicted protein−DNA binding free energy versus experimentally measured ones. The line is from linear regression fit with a correlation coefficient of 0.83. To facilitate comparison, the DFIRE energy function was scaled with a constant scaling factor of 0.0017 and shifted by −5.50 in log K_D units. The dashed line is from the regression fit.

Protein−Protein and Protein−DNA Complexes. The DFIRE trained by the structures of proteins with small ligands (M_W < 1000) can also give an accurate prediction of the binding affinity of protein−protein and protein−DNA complexes. Theoretically predicted and experimentally measured binding affinities for protein−protein and protein−DNA complexes are shown in Figures 5 and 6, respectively. Significant correlations are found in both cases with a correlation coefficient of 0.73 for 82 protein−protein (peptide) complexes and 0.83 for 45 protein−DNA complexes. The results are also tabulated in Tables 5 and 6.

Discussion

In this paper, we extended the DFIRE-based energy originally developed for intra- and interprotein inter-

Table 5. Comparison between Experimentally Measured Binding Free Energies and the Theoretically Predicted Ones

PDB ID[a]	interface[b]	expl[c]	DFIRE[d]	PDB ID[a]	interface[b]	expl[c]	DFIRE[d]
1hbs	ABCD/EFGH	−4.8[42]	−8.89	2tpi	ZI/S	−5.8[43]	−7.09
1tce	A/B	−5.8[44]	−8.44	1ak4	A/C	−6.5[45]	−9.40
1lck	A/B	−7.0[44]	−8.76	1b4z	A/B	−7.1[46]	−8.60
1b46	A/B	−7.2[46]	−8.63	2olb	A/B	−7.6[46]	−8.73
1lcj	A/B	−7.8[44]	−8.54	3tpi	Z/S	−7.8[47]	−7.16
1b3l	A/B	−8.0[46]	−8.13	1qka	A/B	−8.1[46]	−8.88
1b9j	A/B	−8.1[46]	−8.8	2pld	A/B	−9.0[44]	−9.83
1b58	A/B	−9.0[46]	−9.51	1sps	A/D	−9.1[44]	−8.34
1dkz	A/B	−9.1[48]	−9.33	1b3g	A/B	−9.2[46]	−8.76
1jeu	A/B	−9.3[46]	−8.66	1b3f	A/B	−9.4[46]	−8.85
1jev	A/B	−9.4[46]	−9.78	1ola	A/B	−9.5[41]	−8.93
1b5i	A/B	−9.6[46]	−8.73	1b05	A/B	−9.7[46]	−8.55
1b32	A/B	−9.7[46]	−8.86	1b52	A/B	−9.7[46]	−8.60
2er6	E/I	−9.8[41]	−11.32	1jet	A/B	−9.8[46]	−11.50
1b40	A/B	−9.9[46]	−9.27	1b51	A/B	−10.0[46]	−8.42
1qkb	A/B	−10.0[46]	−8.66	1nmb	HL/N	−10.0[49]	−12.09
2pcc	A/B	−10.0[50]	−10.01	1b5j	A/B	−10.1[46]	−8.89
1gua	A/B	−10.1[51]	−10.38	1dkg	AB/D	−10.3[52]	−13.38
1ycs	A/B	−10.3[53]	−11.33	1fdl	HL/Y	−11.4[54]	−12.08
1vfb	AB/C	−11.4[55]	−12.50	2jel	HL/P	−11.5[56]	−12.14
1abi	HL/I	−11.6[57]	−12.57	1ebp	A/C	−11.7[58]	−8.03
4sgb	E/I	−11.7[59]	−11.73	1jhl	HL/A	−11.8[60]	−12.55
1nsn	HL/S	−11.8[61]	−13.31	2kai	AB/I	−12.4[62]	−13.27
2sic	E/I	−12.7[63]	−13.13	3sgb	E/I	−12.7[64]	−10.80
1igc	HL/A	−12.7[65]	−11.50	1hwg	A/C	−13.0[66]	−12.51
1cse	E/I	−13.1[67]	−12.20	3hfm	HL/Y	−13.3[68]	−14.55
1ppf	E/I	−13.4[69]	−12.72	3hhr	A/C	−13.6[70]	−12.34
1tec	E/I	−14.0[67]	−12.71	3hfl	HL/Y	−14.5[43]	−13.11
1cho	E/I	−14.6[71]	−13.08	4htc	HL/I	−15.4[72]	−15.68
2sni	E/I	−15.8[73]	−12.91	3ssi	symmetry[e]	−16.0[74]	−13.07
1acb	E/I	−16.1[75]	−12.43	1bth	HL/P	−16.5[66]	−14.30
1efn	A/B	−16.6[76]	−10.88	1brs	B/E	−17.3[77]	−13.00
1tbq	JK/S	−17.3[78]	−17.79	4tpi	Z/I	−17.7[79]	−12.37
1tpa	E/I	−17.8[43]	−12.40	1dfj	E/I	−18.0[80]	−16.50
2ptc	E/I	−18.1[81]	12.36	1avw	A/B	−12.3[82]	−15.33
1fss	A/B	−14.9[83]	−14.16	1stf	E/I	−13.5[84]	−13.83
1ahw	AB/C	−11.5[85]	−15.96	1mLc	AB/E	−9.7[82]	−12.73
1wej	LH/F	−9.5[82]	−12.03	1bql	LH/Y	−14.5[82]	−12.96
1mel	M/B	−10.5[86]	−13.29	1avz	B/C	−6.4[87]	−10.50
1mda	LH/A	−7.3[88]	−11.58	1a0o	A/B	−8.1[82]	−9.95
1atn	A/D	−11.8[89]	−10.64	1gla	G/F	−6.7[90]	−8.77

[a] The database does not include proteins with metal atoms and other non-amino acid components at the interface. [b] The chain IDs that make the interface. [c] Experimental results (in kcal/mol). [d] Predicted values (in kcal/mol) by the DFIRE energy function scaled by a constant factor of 0.0059 and shifted by −5.35 kcal/mol. [e] The second component of complex 3ssi was generated with the symmetry axis provided by the PDB file.

actions to protein−ligand and protein−DNA interactions. This was achieved by using 19 atom types that not only represent atoms contained in proteins but also those used by ligands and DNA. The resulting energy function is shown to be one of the best energy functions for the prediction of the binding affinities of protein−ligand complexes when compared to 12 other scoring functions for all testing sets used in this study. Moreover, the energy function without any modification can be used directly to predict the binding affinities of protein−protein and protein−DNA complexes with a high accuracy. The comparison made here, however, is based solely on published results. It is possible that unpublished upgrades of the compared methods may perform better than DFIRE.

A correlation coefficient of 0.6 for protein−ligand complexes, however, is far from perfect. It is of interest to analyze the performance of the DFIRE energy function against different ligand types. Because the DFIRE energy function has a cutoff at 15 Å, the cutoff may lead to a poor modeling of long-range charge−charge interactions than the modeling of the interactions between a neutral ligand and a protein. Thus, one might expect that DFIRE would give a better prediction for neutral ligands than for charged ligands. The results for ligands with different net charges (ranging from −4 to +2) are compared in Figure 3 for the 100 high-resolution protein−ligand complexes. The correlation coefficients are 0.60 for 46 neutral ligands, 0.86 for 22 positively charged ligands, 0.59 for 32 negatively charged ligands, and 0.71 for all 54 charged ligands. Within the negatively charged ligands, the correlation coefficients are 0.63 for 12 ligands with −1 charge and 0.77 for 14 ligands with −2 charge. Thus, contrary to our expectation, the DFIRE energy function appears to yield a better correlation for charged ligands. However, the correlation coefficients vary from 0.59 to 0.86, depending on how charged ligands are grouped. Therefore, the difference for different charge types observed here may be due to the small number of protein−ligand complex structures. There also appears that the regression slope for charged ligands is steeper than for neutral ligands. However, the difference is likely smaller than the error associated with regression analysis. Similar results were obtained when dividing ligands in term of their chemical complexity (the number of atom types). Thus, more studies are needed to further identify the source for the

Table 6. Comparison between Experimentally Measured Protein−DNA Binding Free Energies and the Theoretically Predicted Ones

PDB ID	interface[a]	expl[b]	DFIRE[c]	PDB ID	interface[a]	expl[b]	DFIRE[c]
1cma	AB/CD	−5.5[91]	−7.02	1glu	B/DC	−6.0[92]	−6.62
1ckt	A/BC	−6.6[93]	−7.58	1run	A/CF	−6.8[94]	−6.87
1dp7	PQ/DE	−6.8[95]	−6.26	1azp	A/BC	−6.8[96]	−7.11
1tf3	A/EF	−7.11[97]	−8.93	1b69	A/BC	−7.4[98]	−7.57
1ysa	AB/CD	−7.5[99]	−7.75	1dgcd	AC/BD	−7.57[100]	−7.83
1pue	AB/E	−7.64[101]	−7.61	1ais	A/CE	−7.7[102]	−9.22
1oct	AB/C	−7.8[103]	−8.68	1apl	AB/D	−7.82[104]	−7.07
1nk3	AB/P	−7.88[105]	−7.21	1bc7	AB/C	−7.9[106]	−7.58
2gat	A/BC	−8.0[107]	−7.38	1hcr	A/BC	−8.0[108]	−7.65
1tsr	ABC/EF	−8.15[109]	−7.61	1bp7	B/12	−8.15[110]	−9.23
1qrv	A/CD	−8.2[111]	−7.50	1gcc	A/BC	−8.25[112]	−7.55
1yui	A/B	−8.3[113]	−7.42	1mse	AB/C	−8.4[114]	−8.69
1tro	ABCD/IJ	−8.6[115]	−8.35	2dgcd	AC/BD	−8.61[116]	−7.96
1mdy	AB/EF	−8.7[117]	−8.23	1hcq	AB/CD	−8.7[118]	−8.33
1lmb	12/34	−8.73[119]	−8.65	1ytf	ABCD/EF	−8.8[120,121]	−9.64
1aay	A/BC	−8.9[122]	−8.35	1cw0	A/MNO	−9.1[123]	−9.38
1bhm	AB/CD	−9.11[124]	−9.84	1ipp	AB/CD	−9.15[125]	−10.45
1pnrd	AC/BD	−9.3[126]	−9.38	1ihf	AB/CDE	−9.3[127]	−10.09
1cdw	A/BC	−9.32[128]	−9.58	1ecr	A/BC	−10.7[129]	−10.75
1nfk	AB/CD	−10.8[130]	−8.99	1efa	AB/CD	−10.80[131]	−9.58
1par	ABCD/EF	−10.90[132]	−10.50	1a73	ABC/EF	−11.0[133]	−10.48
1l1m	AB/CD	−11.2[134]	−10.47	1az0	AB/CD	−12.9[135]	−10.84
1ca5	A/BC	−6.0[136]	−7.11				

[a] The chain IDs that make the interface. [b] Experimental results (in log K_D units). [c] DFIRE predicted values scaled by a constant factor of 0.0017 and shifted by −5.50 in log K_D units. [d] The complete structures were downloaded from the Nucleic Acid Database (http://ndbserver.rutgers.edu/index.html).

errors in binding prediction so that the accuracy of predicting binding affinity can be further improved.

The 19 atom types (12 for amino acid residues) used in the DFIRE energy function for protein−ligand interaction are a significant reduction from 167 residue-specific atom types for 20 amino acid residues used in the original DFIRE for protein folding and protein−protein binding studies.[27,33] This significantly smaller number of atom types leads to a change of correlation coefficient from 0.79 to 0.73 for 82 protein−protein/peptide complexes. This suggests that an expansion of atom types may be useful to improve somewhat the prediction accuracy of binding affinity. Work in this area is in progress.

One parameter used in this study is the cutoff distance used to define atoms in the binding interface. We found that changing this cutoff distance (9.5 Å used here) does not make a significant change in overall correlation between experimentally measured and theoretically predicted binding affinities. The value 9.5 Å is found to be an optimal cutoff value for the modest success rate in native (or near-native) discrimination. The success rate is 35%, 37%, and 35% at a cutoff distance of 9, 9.5, and 10 Å, respectively. However, it is not an optimal value for the number of complexes with significant correlations. A larger cutoff (for example, 10 Å) increases the number of complexes with higher correlation coefficients (e.g. 23 increases to 26 for $R_s \geq 0.8$, also See Table 4). This result suggests that a higher correlation coefficient between rmsd and energy score does not necessary mean an increase in success rate for ranking native or near native conformation based on energy scores. This perhaps explains why different methods rank so differently in Tables 3 and 4. For example, X-Score is ranked number 2 in Table 4 but number 7 in Table 3 if a near native is defined as an rmsd of ≤1.0 Å. Cerius2/PLP is ranked number 1 in Table 3 but only number 5 in Table 4.

There is one feature that distinguishes the atomic DFIRE energy function from many existing empirical and statistical energy functions. That is, the DFIRE's performance is robust against the database used to train the energy function. For example, the DFIRE energy function for protein−ligand interaction trained with the structures of single-chain proteins can provide similar level of accuracy in predicting binding affinity as the energy function trained by the database of protein−ligand structures (Table 2). This occurs despite the fact that all interactions involving atom types N.ar, S.o2, P.3, F, Cl, Br, and Met are zero in the energy function trained by single-chain proteins only. This database independence is also evident from the fact that the energy function trained by protein−ligand complexes can be directly used for protein−DNA complexes. Thus, the results suggest that the main contribution to the accuracy of the DFIRE energy function is from the atom types contained in proteins. Indeed, we found that atom types N.ar, S.o2, P.3, F, Cl, Br, and Met contribute only 8% of the contacts (at a cutoff distance of 9.5 Å) in the high-resolution 100 protein−ligand complexes. Indeed, upon setting all interactions involving atom types N.ar, S.o2, P.3, F, Cl, Br, and Met to zero in the energy function trained by the 200 protein−ligand complexes, the correlation coefficient increases slightly from 0.63 to 0.64 for the 100 protein−ligand complexes and decreases slightly from 0.63 to 0.62 in the testing set of 30 protein−ligand complexes. However, for specific protein−ligand complexes, interactions involving atom types N.ar, S.o2, P.3, F, Cl, Br, and Met may well be important.

The DFIRE energy function developed here, however, is not fully transferable across different systems. Although the energy function can produce significant correlations for the binding affinities of protein−ligand, protein−protein, and protein−DNA complexes, the regression slope has to be modified from one system to another and even within the different training and testing sets of protein−ligand complexes. This is different from the DFIRE energy function built on 167 atom types for protein folding and protein−protein interactions. The regression slopes between experimental and theoretical results for mutation-induced stability changes and protein−protein (peptide) binding affinities are essentially the same. Thus, increasing the number of atom types from 19 used here may be useful to generate a more uniform regression slope. There also exists a different level of accuracy of the DFIRE energy function for different systems; the correlation coefficients are around 0.6 for different sets of protein−ligand complexes, 0.7 for 82 protein−protein (peptide) complexes, and 0.8 for 45 protein−DNA complexes. We found that the correlation coefficients between experimental binding affinity and the number of atomic contacts at the interface are 0.55 for protein−ligand complexes (the 100 protein−ligand set), 0.66 for 82 protein−protein complexes, but only 0.21 for 45 protein−DNA complexes. The last correlation is due to the existence of several significant outliers (their removal will produce a correlation coefficient of 0.65). Thus, the correlation between the number of interfacial atomic pairs and binding affinities cannot fully account for the different level of accuracy in different systems. One possible reason is that protein−ligand interactions are more complex than those between protein and protein and those between protein and DNA because of the diversity of ligand types.

Acknowledgment. We would like to thank Dr. Renxiao Wang and Prof. Shaomeng Wang for providing us their X-Score program and benchmarks. This work was supported by NIH (R01 GM 966049 and R01 GM 068530), a grant from HHMI to SUNY Buffalo, and by the Center for Computational Research and the Keck Center for Computational Biology at SUNY Buffalo. Y.Z. was also supported by a two-base fund from National Science Foundation of China.

References

(1) Moult, J.; Fidelis, K.; Zemla, A.; Hubbard, T. Critical assessment of methods of protein structure prediction (CASP)—Round V. *Proteins* **2003**, *53*, 334−339.

(2) Krumrine, J.; Raubacher, F.; Brooijmans, N.; Kuntz, I. Principles and methods of docking and ligand design. *Methods Biochem. Anal.* **2003**, *44*, 443−476.

(3) Glen, R. C.; Allen, S. C. Ligand-protein docking: Cancer research at the interface between biology and chemistry. *Curr. Med. Chem.* **2003**, *10*, 763−767.

(4) Jani, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J. CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins* **2003**, *52*, 2−9.

(5) Kono, H.; Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins* **1999**, *35*, 114−131.

(6) Selvaraj, S.; Kono, H.; Sarai, A. Specificity of protein−DNA recognition revealed by structure-based potentials: Symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.* **2002**, *322*, 907−915.

(7) Gromiha, M. M.; Siebers, J. G.; Selvaraj, S.; Kono, H.; Sarai, A. Intermolecular and intramolecular readout mechanisms in protein−DNA recognition. *J. Mol. Biol.* **2004**, *337*, 285−294.

(8) Dill, K. A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133−7155.

(9) Pace, C. N.; Shirley, B. A.; McNutt, M.; Gajiwala, K. Forces contributing to the conformational stability of proteins. *FASEB J.* **1996**, *10*, 75−83.

(10) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(11) Morris, G.; Goodsell, D.; Halliday, R.; Huey, R.; Hart, W.; Belew, R. K.; Olson, A. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(12) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acct. Chem. Res.* **2000**, *33*, 889−897.

(13) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(14) Bohm, H. J. 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.

(15) Gehlhaar, D.; Verkhivker, G.; Rejto, P.; Sherman, C.; Fogel, D.; Fogel, L.; Freer, S. Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *95*, 317−324.

(16) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1999**, *261*, 470−489.

(17) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(18) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11−26.

(19) DeWitte, R.; Shakhnovich, E. 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.

(20) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP—Potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(21) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Foster, M. J.; Thornton, J. M. 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.

(22) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Comput. Chem.* **1999**, *42*, 791−804.

(23) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(24) Ishchenko, A.; Shakhnovich, E. Small molecule growth 2001 (SMOG2001)—an improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.

(25) Jiang, L.; Gao, Y.; Mao, F.; Liu, Z.; Lai, L. Potential of mean force for protein−protein interaction studies. *Proteins* **2002**, *46*, 190−196.

(26) Lu, H.; Lu, L.; Skolnick, J. Development of unified statistical potentials describing protein−protein interactions. *Biophys. J.* **2003**, *84*, 1895−1901.

(27) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins* **2004**, *56*, 93−101.

(28) Thomas, P. D.; Dill, K. A. Statistical potentials extracted from protein structures: How accurate are they? *J. Mol. Biol.* **1996**, *257*, 457−469.

(29) Sippl, M. J. 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.

(30) Furuichi, E.; Koehl, P. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins* **1998**, *31*, 139−149.

(31) Moont, G.; Gabb, H.; Sternberg, M. Use of pair potentials across protein interfaces in screening predicted docked complexes. *Proteins* **1999**, *35*, 364−373.

(32) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Pers. Drug. Discovery Des.* **2000**, *20*, 99−114.

(33) 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. Corrections, **2003**, *12*, 2121.

(34) Zhou, H.; Zhou, Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* **2003**, *54*, 315−322.

(35) Zhou, H.; Zhou, Y. Single-body knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* **2004**, *55*, 1005−1013.

(36) Zhang, C.; Liu, S.; Zhou, H.; Zhou, Y. The dependence of all-atom statistical potentials on training structural database. *Biophys. J.* **2004**, *86*, 3349−3358.

(37) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(38) Wang, G.; Dunbrack, R. L., Jr. PISCES: A protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.

(39) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(40) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, L., Reddy, M. R., Eds.; American Chemical Society: Washington, DC, 1999.

(41) Puvanendrampillai, D.; Mitchell, J. Protein Ligand Database (PLD): Additional understanding of the nature and specificity of protein−ligand complexes. *Bioinformatics* **2003**, *19*, 1856−1857.

(42) Ross, P. D.; Hofrichter, J.; Eaton, W. A. Thermodynamics of gelation of sickle cell deoxyhemoglobin. *J. Mol. Biol.* **1977**, *115*, 111−134.

(43) Horton, N.; Lewis, M. Calculation of the free energy of association for protein complexes. *Protein Sci.* **1992**, *1*, 169−181.

(44) Zhou, Y.; Abagyan, A. R. How and why phosphotyrosine-containing peptides bind to the SH2 and PTB domains. *Fold. Des.* **1998**, *3* (6), 513−522.

(45) Gamble, T. R.; Vajdos, F. F.; Yoo, S.; Worthylake, D. K.; Houseweart, M.; Sundquist, W. I.; Hill, C. P. Crystal structure of human cyclophilin a bound to the amino-terminal domain of HIV-1 capsid. *Cell* **1996**, *87*, 1285−1294.

(46) Wang, T.; Wade, R. C. Comparative binding energy (COMBINE) analysis of oppA-Peptide complexes to relate structure to binding thermodynamics. *J. Med. Chem.* **2002**, *45*, 4828−4837.

(47) Vincent, J. P.; Lazdunski, M. Preexistence of the active site in zymogens, the interaction of trypsinogen with the basic pancreatic trypsin inhibitor (Kunitz). *Febs. Lett.* **1976**, *63* (2), 240−244.

(48) 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.

(49) Tulip, W. R.; Harley, V. R.; Webster, R. G.; 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.

(50) Corin, A. F.; McLendon, G.; Zhang, Q. P.; Hake, R. A.; Falvo, J.; Lu, K. S.; Ciccarelli, R. B.; Holzschu, D. Effects of surface amino acid replacements in cytochrome *c* peroxidase on complex formation with cytochrome *c*. *Biochemistry* **1991**, *30*, 11585−11595.

(51) 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.

(52) Harrison, C. J.; HayerHartl, M.; DiLiberto, M.; Hartl, F. U.; Kuriyan, J. Crystal structure of the nucleotide exchange factor GrpEbound to the ATPase domain of the molecular chaperone DnaK. *Science* **1997**, *276*, 431−435.

(53) Gorina, S.; Pavletich, N. P. Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science* **1996**, *274*, 1001−1005.

(54) Arnon, R. Synthetic peptides as the basis for vaccine design. *Mol. Immunol.* **1991**, *28*, 209−215.

(55) Verhoeyen, M.; Milstein, C.; Winter, G. Reshaping human antibodies: Grafting an antilysozyme activity. *Science* **1988**, *239*, 1534−1536.

(56) Smallshaw, J. E.; Brokx, S.; Lee, J. S.; Waygood, E. B. Determination of the binding constants for three HPr-specific monoclonal antibodies and their fab fragments. *J. Mol. Biol.* **1998**, *325*, 765−774.

(57) Qiu, X.; Padmanabhan, K. P.; Carperos, V. E.; Tulinsky, A.; Kline, T.; Maraganore, J. M.; Fenton, J. W., II. Structure of the hirulog 3-thrombin complex and nature of the S′ subsites of substrates and inhibitors. *Biochemistry* **1992**, *31*, 11689−11697.

(58) Lee, C. H.; Saksela, K.; Mirza, U. A.; Chait, B. T.; Kuriyan, J. Crystal structure of the conserved core of HIV-1 Nef complexed with a src family SH3 domain. *Cell* **1996**, *85*, 931−942.

(59) Hass, G. M.; Ryan, C. A. Carboxypeptidase inhibitor from potatoes. *Methods Enzymol.* **1981**, *80*, 779−790.

(60) Chitarra, V.; Alzari, P. M.; Bentley, G. A.; Bhat, T. N.; Eisele, J.; Houdusse, A.; Lescar, J.; Souchon, H.; Poljak, R. J. Three-dimensional structure of a heteroclitic antigen−antibody cross-reaction complex. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7711−7715.

(61) Smith, A. M.; Benjamin, D. C. The antigenic surface of staphylococcal nuclease. 2. Analysis of the n-1 epitope by site-directed mutagenesis. *J. Immunol.* **1991**, *146*, 1259−1264.

(62) 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.

(63) 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.

(64) Bigler, T. L.; Lu, W.; Park, S. J.; Tashiro, M.; Wieczorek, M.; Wynn, R.; Laskowski, M. J. 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.

(65) Sjobring, U.; Bjorck, L.; Kastern, W. Streptococal protein G gene structure and protein bind properties. *J Biol. Chem.* **1991**, *266*, 399−405.

(66) Guinto, E. R.; Ye, J.; Bonniec, B. F. L.; Esmon, C. T. 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.

(67) Seemüller, U.; Fritz, H.; Euliz, M. Eglin elastase-cathepsing inhibitor from leeches. *Methods Enzymol.* **1981**, *80*, 804−816.

(68) 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. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 5938−5942.

(69) 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 chloromethyl ketone and significance of the TYR−PRO−PRO−TRP insertion segment. *EMBO J.* **1989**, *8*, 3467−3475.

(70) Vos, A. D.; Ultsch, M.; Kossiakoff, A. Human growth hormone and extracellular domain of its receptor: Crystal structure of the complex. *Science* **1992**, *255*, 306−312.

(71) Lu, W. Y.; Qasim, M. A.; Laskowski, M.; Kent, S. B. H. 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.

(72) Ayala, Y.; Vindigni, A.; Nayal, M.; Spolar, R.; Record, M. J.; Cera, E. D. Thermodynamic investigation of hirudin binding to the slow and fast forms of thrombin: Evidence for folding transitions in the inhibitor and protease coupled to binding. *J. Mol. Biol.* **1995**, *253*, 787−798.

(73) Svendsen, I.; Jonassen, I.; Hejgaard, J.; Boisen, S. Amino acid sequence homology between a serine protease inhibitor from barely *hordeum vulgare* cultivar hiproly and potato inhibitor I. *Carsberg Res. Commun.* **1980**, *45*, 389−395.

(74) 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.

(75) Ascenzi, P.; Amiconi, G.; Menegatti, E.; Guarneri, M.; Bolognesi, M.; Schnebli, H. Binding of the recombinant proteinase inhibitor eglin c from leech *Hirudo medicinalis* to human leukocyte elastase, bovine alpha-chymotrypsin and subtilisin carlsberg: Thermodynamic study. *J. Enzyme Inhib.* **1988**, *2*, 167−172.

(76) Clackson, T.; Ultsch, M. H.; Wells, J. A.; de Vos, A. M. 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.

(77) Hartley, R. W. Directed mutagenesis and barnase−barstar recognition. *Biochemistry* **1993**, *32*, 5978−5984.

(78) van de Locht, 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.

(79) Wallqvist, A.; Jernigan, R. L.; Covell, D. G. A preference-based free energy parametrization of enzyme−inhibitor binding applications to HIV-1 protease inhibitor design. *Protein Sci.* **1995**, *4*, 1881−1903.

(80) Kobe, B.; Deisenhofer, J. A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature* **1995**, *374*, 183−186.

(81) Menegatti, E.; Guarneri, M.; Bolognesi, M.; Ascenzi, P.; Amiconi, G. Binding of the bovine pancreatic secretory trypsin inhibitor (kazal) to bovine serine (pro)enzymes. *J. Mol. Biol.* **1987**, *198*, 129−132.

(82) Stites, W. Protein−protein interactions: Interface structure, binding thermodynamics, and mutational analysis. *Chem. Rev.* **1997**, *97*, 1233−1250.

(83) Eastman, J.; Wilson, E.; Cerveansky, C.; Rosenberry, T. Fasciculin 2 binds to the peripheral site on acetylcholinesterase and inhibits substrate hydrolysis by slowing a step involving proton transfer during enzyme acylation. *J. Biol. Chem.* **1995**, *270*, 19694−19701.

(84) Turk, B.; Krizaj, I.; Kralj, B.; Dolenc, I.; Popovic, T.; Bieth, J.; Turk, V. Bovine stefin c, a new member of the stefin family. *J. Biol. Chem.* **1993**, *268*, 7323−7329.

(85) Huang, M.; Syed, R.; Stura, E.; Stone, M.; Stefanko, R.; Ruf, W.; Edgington, T.; Wilson, I. The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF. G9 complex. *J. Mol. Biol.* **1998**, *275*, 873−894.

(86) Ward, E.; Gussow, D.; Griffiths, A.; Jones, P.; Winter, G. Binding activities of a repertoire of single immunoglobulin variable domains secreted from *Escherichia coli*. *Nature* **1989**, *341*, 544−546.

(87) Lee, C.; Leung, B.; Lemmon, M.; Zheng, J.; Cowburn, D.; Kuriyan, J.; Saksela, K. A single amino acid in the SH3 domain of Hck determines its high affinity and specificity in binding to HIV-1 Nef protein. *EMBO J.* **1995**, *14*, 5006−5015.

(88) Davidson, V.; Graichen, M.; Jones, L. Binding constants for a physiologic electron-transfer protein complex between methylamine dehydrogenase and amicyanin. effects of ionic strength and bound copper on binding. *Biochim. Biophys. Acta* **1993**, *1144*, 39−45.

(89) Mannherz, H.; Goody, R.; Konrad, M.; Nowak, E. The interaction of bovine pancreatic deoxyribonuclease I and skeletal muscle actin. *Eur. J. Biochem.* **1980**, *104*, 367−379.

(90) Pettigrew, D.; Meadow, N.; Roseman, S.; Remington, S. Cation-promoted association of escherichia coli phosphocarrier protein IIAGlc with regulatory target protein glycerol kinase: Substitutions of a zinc(II) ligand and implications for inducer exclusion. *Biochemistry* **1998**, *37*, 4875−4883.

(91) Hyre, D.; Spicer, L. Thermodynamic evaluation of binding interactions in the methionine repressor system of escherichia-coli using isothermal titration calorimetry. *Biochemistry* **1995**, *34*, 3212−3221.

(92) Lundback, T.; Cairns, C.; Gustafsson, J.; Carlstedtduke, J.; Hard, T. Thermodynamics of the glucocorticoid receptor-DNA interaction: Binding of wild-type GR DBD to different response elements. *Biochemistry* **1993**, *32*, 5074−5082.

(93) Jamieson, E. R.; Lippard, S. J. Stopped-flow fluorescence studies of HMG-domain protein binding to cisplatin-modified DNA. *Biochemistry* **2000**, *39*, 8426−8438.

(94) Shi, Y.; Wang, S.; Krueger, S.; Schwarz, F. Effect of mutations at the monomer-monomer interface of cAMP receptor protein on specific DNA binding. *J. Biol. Chem.* **1999**, *274*, 6946−6956.

(95) Cornille, F.; Emery, P.; Schuler, W.; Lenoir, C.; Mach, B.; Rogues, B.; Reith, W. DNA binding properties of a chemically synthesized DNA binding domain of hRFX1. *Nucleic Acids Res.* **1998**, *26*, 2143−2149.

(96) McAfee, J.; Edmondson, S.; Zegar, I.; Shriver, J. Equilibrium DNA binding of Sac7d protein from the hyperthermophile sulfolobus acidocaldarius: Fluorescence and circular dichroism studies. *Biochemistry* **1996**, *35*, 4034−4045.

(97) Liggins, J.; Privalov, P. Energetics of the specific binding interaction of the first three zinc fingers of the transcription factor TFIIIA with its cognate DNA sequence. *Proteins* **2000**, *Suppl. 4*, 50.

(98) Connolly, K.; Ilangovan, U.; Wojciak, J.; Iwahara, M.; Clubb, R. Major groove recognition by three-stranded β-sheets: Affinity determinants and conserved structural features. *J. Mol. Biol.* **2000**, *300*, 841−856.

(99) Hollenbeck, J.; McClain, D.; Oakley, M. The role of helix stabilizing residues in GCN4 basic region folding and DNA binding. *Protein Science* **2002**, *11*, 2740−2747.

(100) Berger, C.; Jelesarov, I.; Bosshard, H. Coupled folding and site-specific binding of the GCN4-bZIP transcription factor to the AP-1 and ATF/CREB DNA sites studied by microcalorimetry. *Biochemistry* **1996**, *35*, 14984−14991.

(101) Gross, P.; Yee, A.; Arrowsmith, C.; Macgregor, R. Quantitative hydroxyl radical footprinting reveals cooperative interactions between DNA-binding subdomains of PU.1 and IRF4. *Biochemistry* **1998**, *37*, 9802−9811.

(102) O'Brien, R.; DeDecker, B.; Fleming, K.; Sigler, P.; Ladbury, J. The effects of salt on the TATA binding protein−DNA interaction from a hyperthermophilic archaeon. *J. Mol. Biol.* **1998**, *279*, 117−125.

(103) Lundback, T.; Chang, J.; Phillips, K.; Luisi, B.; Ladbury, J. Characterization of sequence-specific DNA binding by the transcription factor Oct-1. *Biochemistry* **2000**, *39*, 7570−7579.

(104) Carra, J.; Privalov, P. Energetics of folding and DNA binding of the MATα2 homeodomain. *Biochemistry* **1997**, *36*, 526−535.

(105) Koizumi, K.; Lintas, C.; Nirenberg, M.; Maeng, J.; Ju, J.; Mack, J.; Gruschus, J.; Odenwald, W.; Ferretti, J. Mutations that affect the ability of the vnd/NK-2 homeoprotein to regulate gene expression: Transgenic alterations and tertiary structure. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3119−3124.

(106) Szymczyna, B.; Arrowsmith, C. DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein−DNA recognition. *J. Biol. Chem.* **2000**, *275*, 28363−28370.

(107) Foti, M.; Omichinski, J.; Stahl, S.; Maloney, D.; West, J.; Schweitzer, B. Effects of nucleoside analogue incorporation on DNA binding to the DNA binding domain of the GATA-1 erythroid transcription factor. *FEBS Lett.* **1999**, *444*, 47−53.

(108) Robinson, C.; Sligar, S. Participation of water in hin recombinase-DNA recognition. *Protein Sci.* **1996**, *5*, 2119−2124.

(109) Nagaich, A. K.; Apella, E.; Harrington, R. E. DNA bending is essential for the site-specific recognition of DNA response elements by the DNA binding domain of the tumor suppressor protein p53. *J. Biol. Chem.* **1997**, *272*, 14842−14849.

(110) Seligman, L.; Chevalier, B.; Chadsey, M.; Edwards, S.; Savage, J.; A.L. V. Mutations altering the cleavage specificity of a homing endonuclease. *Nucleic Acids Res.* **2002**, *30*, 3870−3879.

(111) Klass, J.; Murphy, F.; Fouts, S.; Serenil, M.; Changela, A.; Siple, J.; Churchill, M. The role of intercalating residues in chromosomal high-mobility-group protein DNA binding, bending and specificity. *Nucleic Acids Res.* **2003**, *31*, 2852−2864.

(112) Hao, D.; Yamasaki, K.; Sarai, A.; Ohme-Takagi, M. Determinants in the sequence specific binding of two plant transcription factors, CBF1 and NtERF2, to the DRE and GCC motifs. *Biochemistry* **2003**, *41*, 4202−4208.

(113) Pedone, P.; Ghirlando, R.; Clore, G.; Gronenborn, A.; Felsenfeld, G.; Omichinski, J. The single Cys(2)-His(2) zinc finger domain of the GAGA protein flanked by basic residues is sufficient for high-affinity specific DNA binding. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2822−2826.

(114) Tanikawa, J.; Yasukawa, T.; Enari, M.; Ogata, K.; Nishimura, Y.; Ishii, S.; Sarai, A. Recognition of specific DNA sequences by the c-myb protooncogene product: Role of three repeat units in the DNA-binding domain. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9320−9324.

(115) Grillo, A.; Brown, M.; Royer, C. Probind the physical basis for trp repressor-operator recognition. *J. Mol. Biol.* **1999**, *287*, 539−554.

(116) Palmer, C.; Gegnas, L.; A., S. Mechanism of DNA binding enhancement by hepatitis B virus protein pX. *Biochemistry* **1997**, *36*, 15349−15355.

(117) Kunne, A.; Sieber, M.; Meierhans, D.; Allemann, R. Thermodynamics of the DNA binding reaction of transcription factor MASH-1. *Biochemistry* **1998**, *37*, 4271−4223.

(118) Ozers, M.; Hill, J.; Ervin, K.; Wood, J.; Nardulli, A.; Royer, C.; J., G. Equilibrium binding of estrogen receptor with DNA using fluorescence anisotropy. *J. Biol. Chem.* **1997**, *272*, 30405−30411.

(119) Merabet, E.; Ackers, G. K. Calorimetric analysis of λ cI repressor binding to DNA operator sites. *Biochemistry* **1995**, *34*, 8554−8563.

(120) Prabakaran, P.; An, J.; Gromiha, M.; Selvaraj, S.; Uedaira, H.; Kono, H.; Sarai, A. Thermodynamic database for protein-nucleic acid interactions (ProNIT). *Bioinformatics* **2001**, *17*, 1027−1034.

(121) Sarai, A.; Gromiha, M.; An, J.; Prabakaran, P.; Selvaraj, S.; Kono, H.; Oobatake, M.; Uedaira, H. Thermodynamic databases for proteins and protein-nucleic acid interactions. *Biopolymers* **2002**, *61*, 121−126.

(122) Hamilton, T.; Borel, F.; Romaniuk, P. Comparison of the DNA binding characteristics of the related zinc finger proteins WT1 and EGR1. *Biochemistry* **1998**, *37*, 2051−2058.

(123) Gonzalez-Nicieza, R.; Turner, D.; Connolly, B. DNA binding and cleavage selectivity of the escherichia coli DNA G:T-mismatch endonuclease (vsr protein). *J. Mol. Biol.* **2001**, *310*, 501−508.

(124) Engler, L.; Sapienza, P.; Dorner, L.; Kucera, R.; Schildkraut, I.; Jen-Jacobson, L. The energetics of the interaction of BamHI endonuclease with its recognition site GGATCC. *J. Mol. Biol.* **2001**, *307*, 619−636.

(125) Wittmayer, P. K.; Raines, R. T. Substrate binding and turnover by the highly specific I−PpoI endonuclease. *Biochemistry* **1996**, *35*, 1076−1083.

(126) Moraitis, M. I.; Xu, H.; Matthews, K. S. Ion concentration and temperature dependence of DNA binding: Comparison of PurR and LacI repressor proteins *Biochemistry* **2001**, *40*, 8109−8117.

(127) Murtin, C.; Engelhorn, M.; Geiselmann, J.; Boccard, F. A quantitative UV laser footprinting analysis of the interaction of IHF with specific binding sites: Re-evaluation of the effective concentration of IHF in the cell. *J. Mol. Biol.* **1998**, *284*, 949−961.

(128) Cohen, S.; Jamieson, E.; Lippard, S. Enhanced binding of the TATA-binding protein to TATA boxes containing flanking cisplatin 1,2-*cross*-links. *Biochemistry* **2000**, *39*, 8259−8265.

(129) Neylon, C.; Brown, S.; Kralicek, A.; Miles, C.; Love, C.; Dixon, N. Interaction of the escherichia coli replication terminator protein (Tus) with DNA: A model derived from DNA-binding studies of mutant proteins by surface plasmon resonance. *Biochemistry* **2000**, *39*, 11989−11999.

(130) Hart, D.; Speight, R.; Cooper, M.; Sutherland, J.; Blackburn, J. The salt dependence of DNA recognition by NF-κ B p50: A detailed kinetic analysis of the effects on affinity and specificity. *Nucleic Acids Res.* **1999**, *27*, 1063−1069.

(131) Frank, D.; Saecker, R.; Bond, J.; Capp, M.; O. V., T.; Melcher, S.; M.M., L.; M.T., R. Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: Effects of converting a consensus site to a nonspecific site. *J. Mol. Biol.* **1997**, *267*, 1186−1206.

(132) Schildbach, J.; Karzai, A.; Raumann, B.; Sauer, R. Origins of DNA-binding specificity: Role of protein contacts with the DNA backbone. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 811−817.

(133) Galburt, E.; Chadsey, M.; Jurica, M.; Chevalier, B.; Erho, D.; Tang, W.; Monnat, R.; Stoddard, B. Conformational changes and cleavage by the homing endonuclease I−Ppol: A critical role for a leucine residue in the active site. *J. Mol. Biol.* **2000**, *300*, 877−887.

(134) Barry, J.; Matthews, K. Substitutions at histidine 74 and aspartate 278 alter ligand binding and allostery in lactose repressor protein. *Biochemistry* **1999**, *38*, 3579−3590.

(135) Martin, A.; Sam, M.; Reich, N.; J.J., P. Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nat. Struct. Biol.* **1999**, *6*, 269−277.

(136) Lundbäck, T.; Hansson, H.; Knapp, S.; Ladenstein, R.; Härd, T. Thermodynamic characterization of non-sequence-specific DNA-binding by the Sso7d protein from *Sulfolobus solfataricus*. *J. Mol. Biol.* **1998**, *276*, 775−786.

JM049314D