M-Score: A Knowledge-Based Potential Scoring Function Accounting for Protein Atom Mobility

Chao-Yie Yang, Renxiao Wang, and Shaomeng Wang*

Departments of Internal Medicine, Pharmacology and Medicinal Chemistry, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109-0934

Received January 15, 2005

A knowledge-based potential scoring function, named M-Score, has been developed based upon 2331 high-resolution crystal structures of protein−ligand complexes. M-Score considers the mobility of protein atoms, describing the location of each protein atom by a Gaussian distribution instead of a fixed position based upon the isotropic *B*-factors. This leads to an increase in the number of atom-pairs in the construction of knowledge-based potentials and a smoothing effect on the pairwise distribution functions. M-Score was validated using 896 complexes which were not included in the 2331 data set and whose experimentally determined binding affinities were available. The overall linear correlation coefficient (*r*) between the calculated scores and experimentally determined binding affinities (pK_i or pK_d) for these 896 complexes is −0.49. Evaluation of M-Score against 17 protein families showed that we obtained good to excellent correlations for six protein families, modest correlations for four protein families, and poor correlations for the remaining seven protein families.

Introduction

Scoring functions play an important role in virtual database screening aimed at the discovery of lead molecules and in structure-based lead optimization.[1,2] In current scoring function development, the following three different approaches are commonly used: force field-based functions,[3,4] empirical functions,[5,6] and knowledge-based potential functions.[7−13] Force field-based scoring functions take advantage of atomic force-field parameters developed for molecular mechanics calculations and molecular dynamics simulation.[3,4] Empirical scoring functions divide the binding free energy into several physical terms and determine the weight of each term by regression analysis using a training set containing experimentally determined binding affinity data of protein−ligand complexes.[5,6] Knowledge-based scoring functions have recently emerged as a new approach,[7−13] in which statistical atomic distances between protein atoms and ligand atoms are used to calculate atom pair distributions. When compared to a reference state using a Boltzman form equation, a "free energy" term is derived to describe each atom-pair interaction. The summation of all atom-pair interactions is then used to assess the overall strength of interaction for each protein−ligand complex. Several such scoring functions have been developed, which include PMF,[7] BLEEP,[8,9] Drugscore,[10] SMoG,[11,12] and DFIRE.[13]

One unique aspect in the development of knowledge-based scoring functions is that they do not rely on regression to retrofit experimentally determined binding affinity data, in contrast to force field-based and empirical scoring functions. In addition, the many-body interactions in protein−ligand complexes, which are difficult to describe in traditional force field-based scoring functions, are embedded in the structural data used to generate the knowledge-based potential functions. Major deficiencies of current knowledge-based scoring functions include an insufficient number of complex structures used to derive atom pair distributions and the lack of a well-defined reference state. A common major shortcoming in all the current scoring functions published to date is that only a single conformation of a protein is used, i.e., the protein structure is treated as a rigid body and protein flexibility is ignored.

In our continuing efforts to construct the PDBbind database,[14,15] we have recently compiled from the Protein Data Bank (PDB)[16] a set of 3227 protein−ligand complex structures whose resolutions are better than 2.5 Å. Motivated by the availability of this large data set of high-resolution experimentally determined crystal structures, we have now developed a new knowledge-based potential scoring function with better statistics for the derived atom-pair distributions and with an attempt to account for protein atom mobility to overcome some of the major deficiencies in the current knowledge-based scoring functions.

In crystal structure determination, protein atom mobility may be modeled via the atomic *B*-factor (or the temperature factor), which includes thermal fluctuation and uncertainty of fitting one average structure to multiple conformers. Although a term called "atomic displacement parameters" has been used to describe more accurately other structural information incorporated into the *B*-factor, an analytical model is used to fit atoms into the electron density maps where each atom is assumed to move harmonically.[17,18] The probability of observing an atom away from its most probable position is approximated by a Gaussian distribution as determined by the *B*-factor. Upon the basis of the atomic *B*-factor data available from crystal structures in the PDB, we can incorporate the probabilistic positions of a protein atom away from its mean position directly into our knowledge-based scoring function development, which is consistent with the statistical nature of the knowledge-based potential scoring function.

In this study, we divided the data set of 3227 structures into two sets. The first set consists of 2331 protein−ligand complex structures, whose experimental binding affinities were not available. The first data set was used to develop a new knowledge-based potential scoring function, which we named M-Score, into which the mobility of protein atoms is incorporated. In addition, a new atom typing scheme distinguishing the side chain and main chain atoms of the proteins was developed and used, and three atom typing schemes for ligands were evaluated. The second set contains 896 protein−ligand

* To whom correspondence should be addressed. Phone: 734-615-0362. Fax: 734-647-9647. E-mail: shaomeng@umich.edu.

10.1021/jm050043w CCC: $33.50
Published on Web 12/09/2005

Table 1. New Atom Types for the Protein Atoms Used in This Work

	atom type	description
1	Cα or Ca	α carbon
2	C(bk)	backbone carbonyl carbon
3	Cβ or Cb	β carbon
4	C.3(sch)	side chain sp3 carbon
5	C.2(N,Q)	side chain sp2 carbon (N,Q)
6	C.2(E,D)	side chain sp2 carbon (E.D)
7	C.cat(R)	charged carbon in R
8	C.ar	carbon in aromatic system
9	N.am	backbon amidic nitrogen
10	N.ar(W,H)	nitrogen in aromatic system (W,H)
11	Npolar(N,Q)	polar and uncharged nitrogen (N,Q)
12	N.4(R,K)	charged nitrogen atom (R,K)
13	O(bk)	backbone carbonyl oxygen
14	O.co2	carboxylic charged oxygen
15	O(polar)	polar and not charged oxygen
16	O.2(N,Q)	carboxylic uncharged oxygen (N,Q)
17	S(M)	sulfur atom in Met
18	S(C)	sulfur atom in Cys
19	Metal	metal ions

complexes with experimentally determined binding affinity data and was used for validation of M-Score.

Methods

The Data Set. A total of 3227 protein−ligand complexes determined by the X-ray crystallography obtained from our PDBbind database[14,15] were used in this work. The selection criteria were reported recently.[15] Briefly, these 3227 complexes meet the following criteria: (a) there is no covalent bond formed between protein and ligand; (b) no more than one ligand is bound to the binding site of the protein; (c) ligands contain only common organic elements (i.e. H, C, O, N, S, P and halogens) and have a molecular weight of less than 1000 or less than 10 amino acid residues; (d) the resolution of each complex is better than 2.5Å. Of these 3227 complexes, experimentally measured K_i or K_d values of 896 complexes had been retrieved from the literature[15] and were used as the validation data set. The structures of the remaining 2331 complexes were used for the development of M-Score.

We used p*K* to denote experimentally determined binding affinities (−log K_i or -log K_d) and did not attempt to make a distinction between them. K_d is a direct measure of the binding affinity between a protein and a ligand, whereas K_i is calculated from the IC_{50} value determined in a competitive binding experiment based on a competitive kinetic model. Because M-Score is not regression-based, these K_i or K_d values would not affect the knowledge-based potentials in the scoring function nor the predictive power of M-Score. Experimentally determined K_i or K_d values for these 896 complexes were only used in the evaluation of M-Score.

Determination of the protonation state of atoms requires careful experimental studies, and such information was not always available when binding affinity was reported. Therefore, in both scoring function construction and validation, protonation states of atoms in proteins and ligands are assigned according to physiological pH value of 7.4.

Atom Typing. A new atom typing scheme was developed for protein atoms (Table 1). This new atom typing scheme considers both bond orders and atomic polarity, and distinguishes between side chain and backbone atoms. A total of 19 different atom types are defined for protein atoms. For ligand atoms, we used the Tripos Mol2 atom types[19] and a total of 21 atom types are used in our study for ligand atoms (Table 2). We have also evaluated the two atom typing schemes used in Drugscore[10] and in the PMF scoring function by Muegge and Martin[7] for comparison purposes.

Atomic Mobility Based upon the Gaussian Approximation. The atomic *B*-factors obtained from the crystal structure determination correspond to the magnitude of the thermal atomic motion in crystals. When the structure is determined with very high resolution, the anisotropic motion of each atom can be resolved through refinement procedures to reveal specific directions and magnitudes of the atomic motion. Because the determination of

Table 2. Numbers of Atom Pairs Derived from 2331 Potein−Ligand Complexes Using the Atom Types Defined in Table 1 for Proteins and Tripos Mol2 Atom Types for Ligands with a Cutoff Distance of 10 Å

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
		Cα	C(bk)	Cβ	C.3 (sch)	C.2 (N,Q)	C.2 (E,D)	C.cat (R)	C.ar	N.am	N.ar (W,H)	Npolar (N,Q)	N.4 (R,K)	O(bk)	O.co2	O (polar)	O.2 (N,Q)	S(M)	S(C)	Metal
1	C.3	309986	304224	283402	316104	25321	14714	17233	364784	313306	39050	42899	44898	294438	80654	64392	49938	8580	6274	3845
2	C.2	112019	109914	102785	119617	8860	4820	6530	118038	112684	12308	15611	17056	105558	25511	23475	16725	3204	2895	1212
3	C.ar	165261	164149	143753	165460	10431	5458	5808	154258	168235	17608	16187	15538	158721	31158	29273	19907	4801	5301	1348
4	C.1	422	430	368	392	0	0	0	540	435	0	0	0	409	69	54	0	0	0	0
5	C.cat	4176	3982	3408	3068	271	124	104	3590	4068	366	373	227	3876	839	812	502	0	280	0
6	N.3	4044	3859	3695	4236	337	220	214	4800	3999	593	556	584	3737	1073	887	649	128	60	0
7	N.am	39296	38330	36438	42326	3181	1642	2395	45932	39324	4961	5609	6030	36788	10061	8859	6431	796	851	408
8	N.2	14407	14068	12807	13717	1128	604	571	13211	14772	1622	1721	1665	13516	3159	2849	2060	420	474	159
9	N.ar	14059	13789	12175	14057	856	591	447	10613	14440	1097	1292	1328	13133	2549	2268	1497	433	254	56
10	N.1	205	207	189	250	0	0	0	344	225	0	0	0	205	0	0	0	0	0	0
11	N.4	10553	10321	9766	10486	908	676	557	11024	10568	1086	1503	1460	10197	2979	2548	1733	386	309	101
12	N.pl3	27604	27518	23784	24679	1937	1093	958	23296	28469	2302	2819	2572	26027	5930	4942	3651	594	1177	236
13	O.3	114458	111456	104581	108066	11073	7125	7675	133397	115686	17586	18982	20541	105936	35973	23321	21398	2845	1971	2026
14	O.2	46186	44964	42743	50049	3821	2093	2450	54872	46628	5947	6326	6453	43492	11157	9954	7277	1057	984	497
15	O.co2	87312	85115	80154	90009	7785	4540	7256	75767	87618	9811	15049	19712	81156	22397	20292	14164	1835	1766	1655
16	S.3,S.2,P.3	22999	22376	20668	22542	1847	1123	1741	20958	22989	2780	3630	4819	20881	6234	5286	3748	444	418	555
17	S.o, S.o2	1930	1882	1786	2034	204	116	102	2285	1942	436	318	281	1839	408	389	279	0	0	0
18	F	2881	2857	2639	3228	163	82	91	2652	2924	254	264	256	2811	515	512	320	74	118	0
19	Cl	1254	1260	1113	1354	60	0	0	1088	1286	87	86	87	1217	164	192	120	0	0	0
20	Br	615	601	525	654	0	0	0	502	641	0	0	52	576	100	120	54	0	0	0
21	I	686	662	607	819	0	0	0	508	692	0	0	56	639	72	142	0	0	0	0

the anisotropic *B*-factor requires both additional human effort in the structure refinement procedures and very high quality of crystals, it is not commonly adopted in the X-ray structure determination. Therefore, we used the isotropic *B*-factor available in a typical PDB file in our implementation, where each atom is assumed to vibrate isotropically.

The *B*-factors determined from the electron density map from X-ray diffraction assume that each atom vibrates harmonically around an average position in space. The harmonic approximation leads to a Gaussian probability distribution of observing the atom in space with the *B*-factor defined as $B = 8\pi^2 \langle u^2 \rangle$ and the Gaussian distribution function defined as[19]

$$p(u) = (2\pi \langle u^2 \rangle)^{-3/2} \exp\left(-\frac{1}{2}\frac{u^2}{\langle u^2 \rangle}\right) = \left(\frac{B}{4\pi}\right)^{-3/2} \exp\left(-\frac{1}{2}\frac{8\pi^2 u^2}{B}\right) = (2\pi\sigma^2)^{-3/2} \exp\left(-\frac{1}{2}\frac{u^2}{\sigma^2}\right) \quad (1)$$

where $\langle u^2 \rangle$ is the mean square displacement in one-dimension of the atomic position and $\sigma = \sqrt{B/8\pi^2}$ is the width of the Gaussian distribution. Using eq 1, we can assign the probability of observing an atom around its mean position due to the thermal motion by the *B*-factor and incorporate it into our knowledge-based potential scoring function. The difference between our implementation and the conventional knowledge-based method is that we include the spatial distribution of each atom around the mean position based on the *B*-factor characterized by a Gaussian distribution in real space, whereas conventional knowledge-based potential includes a single position for each protein atom in space with 100% probability.

In our implementation, five Gauss-Legendre quadrature points[20] within a range of $\pm\sigma$ from the atomic mean position are used to represent the probability distribution of each atom and the sum of total probability accounts for 0.895. Although probability of evenly spaced points can be used, the Gauss-Legendre quadrature was shown to approximate the probability density more accurately than the same number of evenly spaced points.[21]

Inclusion of the Atomic Mobility of Proteins in Knowledge-Based Potential. We modified the knowledge-based potential function used in Drugscore[10] to include the atomic mobility of proteins. A normalized distance-dependent pair distribution function, $g_{i,j}(r)$, between atom type *i* and atom type *j* is calculated as follows

$$g_{i,j}(r) = \frac{N^{w}_{i,j}(r)/4\pi r^2}{\Sigma_{k,l} N^{w}_{k,l}(r)/4\pi r^2} \quad (2)$$

where $N^{w}_{i,j}$ is the weighted occurrence with a distance from *r* to *r*+d*r* between atom pairs *i* and *j* and is defined as a sum of weighted delta functions:

$$N^{w}_{i,j}(r) = \Sigma_{\mu}\Sigma_{\nu}[\sum_{n} p_{\mu}(u_n)\cdot\delta(\vec{r}^{\,i}_{\mu} - \vec{r}^{\,j}_{\nu}| + u_n - r)] \quad (3)$$

where r^{i}_{μ} is the mean position of protein atom μ (type i), u_n is the deviation of protein atom μ from its mean position, *n* is the number of atomic position of protein atom μ included, $p_{\mu}(u_n)$ is the probability of observing protein atom μ at u_n according to eq 1, and r^{j}_{ν} is the mean position of ligand atom ν (type *j*).

Although the same probability function can be applied for ligand atoms, multiplication of positional probability of a protein atom with that of a ligand atom will lead to unrealistically close distances between protein and ligand atoms and clashes. To avoid such a situation, only positional probability of protein atoms is considered in the current implementation.

Here, the bin size of d*r* equal to 0.1 Å is used, and the Boltzman-like scoring function (F_{SC}) is calculated as follows:

$$F_{SC} = -\ln\frac{g_{i,j}(r)}{g(r)},\ g(r) = \frac{\Sigma_{i,j} g_{i,j}(r)}{i*j} \quad (4)$$

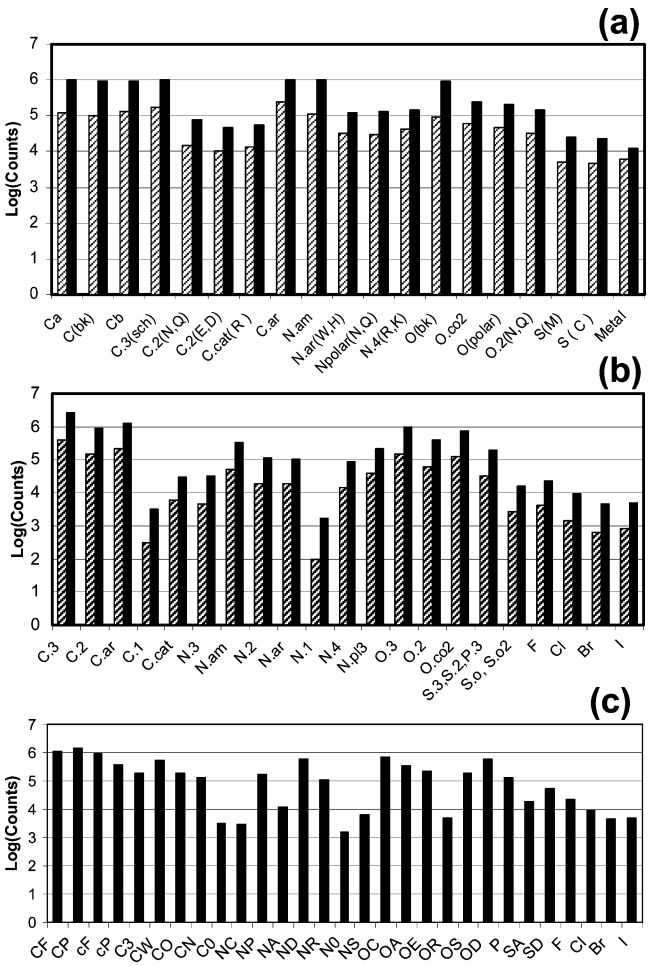

Figure 1. Total counts of distances in atom pairs for (a) each protein atom and (b, c) each ligand atom using cutoff distances of 6 (shaded bars) and 10 Å (solid bars). The atom type for the protein is defined in Table 1. The atom types for the ligands in b and c are the atom types of Mol2 and PMF.

where $g(r)$ is the average of $g_{i,j}(r)$ [defined in eq 2] over pairs of all different atom types separated by a distance of *r*. The cutoff distance for calculating $g_{i,j}(r)$ and the scores is a variable. We have tested cutoff distances of 6, 10, and 12 Å between protein and ligand atoms in our current study. Only non-hydrogen atoms are considered, and metal ions are treated as part of the protein. If the total number of occurrences of an atom pair between proteins and ligands from the whole data set is less than 100, the interaction of this particular atom pair is considered to be statistically insignificant and the corresponding scoring potential is set to zero.

Results and Discussion

Statistics of Atom Pairs Derived from the 2331 Protein−Ligand Complexes. The number of atom pairs derived from these 2331 protein−ligand complexes depends on the atom typing scheme and the cutoff distance, both of which will affect the derived potentials. An ideal design for an atom typing scheme should result in a statistically significant number for each atom pair. Proteins only contain 20 different amino acids, and there are a limited number of atom types. In contrast, small-molecule ligands have very diverse chemical structures and in theory contain far more atom types than are in proteins. In Figure 1, we display the number of atom pairs in our scoring function construction using the atom typing scheme defined in Table 1 for proteins and either the Tripos Mol2 atom types[19] or the atom types of PMF[7] for ligands. Results using either the Mol2 atom

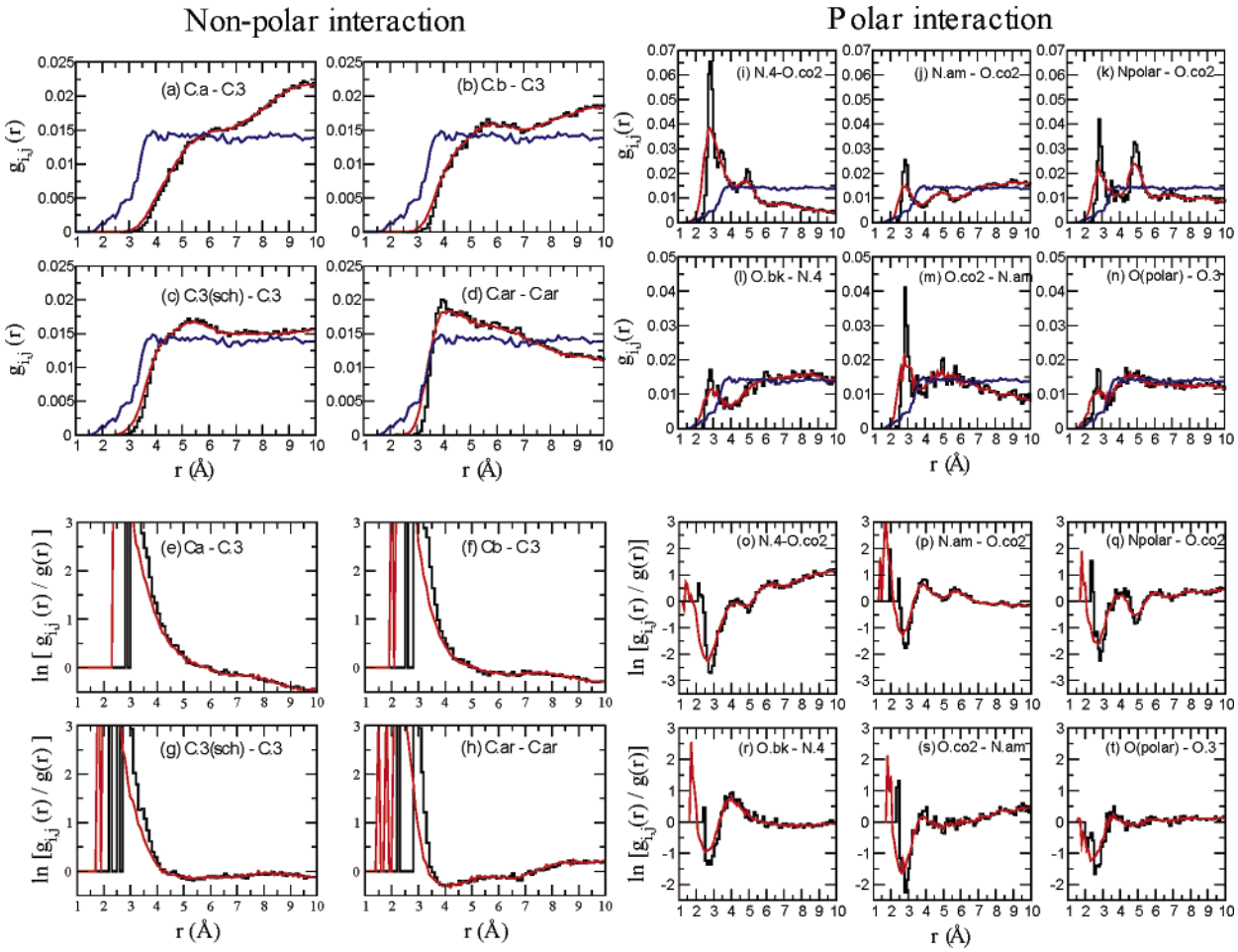

Figure 2. Examples of pairwise distribution functions and knowledge-based free energy potential functions for nonpolar and polar interaction using Tripos Mol2 atom types for ligands. Black line: Rigid protein model; red line: *B*-factor augmentation; blue line: reference state.

types or the atom types of PMF[7] for both proteins and ligands are provided in the Supporting Information.

If we use 6 Å as the cutoff distance for calculation of atom pairs between proteins and ligands, each protein atom type has more than 1000 atom pairs with ligand atoms, with the highest number being 100 000 (shaded bars in Figure 1a). As expected, there are fewer atom pairs that involve either a sulfur atom or metal ions in proteins. When the cutoff distance is set to 10 Å, the number for each atom pair is increased. However, the relative percentage increases in the atom pairs involving backbone protein atoms are somewhat higher than those involving protein side chain atoms and metal ions.

For ligand atom types, we found that the numbers of atom pairs involving C.1, C.cat, N.3, N.1, S.o, S.o2, and halogen atoms in ligands are lower than the numbers of atom pairs involving other atom types by 1−2 log units (Figure 1b). When the cutoff distance is 6 Å, the numbers of two atom pairs involving C.1 and N.1 atom types in ligands are fewer than 1000. When the cutoff distance is set to 10 Å, the numbers for these two atom pairs improve to greater than 1000, although they are still significantly smaller that those for other atom pairs.

In the scoring potential function calculation, the counts for each atom pair determine the pairwise distribution function. When using the Tripos Mol2 atom types for ligands with a cutoff distance of 10 Å, and our atom typing scheme for proteins, the numbers are greater than 10 000 for 119 atom pairs between protein and ligand atoms, between 1000 and 10 000 for 141 atom pairs, below 1000 for the remaining 139 different atom pairs (Table 2). The relative percentages are 30, 35, and 35%, respectively, out of a total of 399 atom pairs between proteins and ligands. In comparison, when the 29 atom types of PMF are used for ligands, the counts are greater than 10 000 for 154 atom pairs between protein and ligand atoms, between 1000 and 10 000 for 170 atom pairs, below 1000 for 227 different atom pairs (Table S1). Their relative percentages are 28, 31, and 41%, out of a total of 551 atom pairs between proteins and ligands. By comparing these two atom typing schemes, the PMF atom typing yields a higher percentage of atom pairs with the number of counts less than 1000. Atom pairs, whose counts are less than 1000, will be referred to as rare atom pairs.

Analysis of the Pairwise Distribution Functions and Potentials. A difference between our new atom typing scheme for proteins and those used previously is the distinction between side chain and backbone atoms. To evaluate if this distinction is important, we have analyzed the pairwise distribution functions and potentials obtained from these 2331 protein−ligand complexes. To exclude the potential influence of the atomic mobility, we generated the pairwise distribution functions and potentials without inclusion of the atomic *B*-factors.

It was found that pairwise distribution functions between Cα, Cβ, and C.3(sch) (all are sp^3 carbon atoms for proteins) and C.3 from ligands have different shapes (Figure 2). The differences among them are more pronounced when the atomic separation is greater than 7 Å. With respect to the potentials for these atom pairs, there is a clear, albeit shallow, attractive potential well between C.3(sch) and C.3, which is not observed between Cα and C.3, and between Cβ and C.3. This may be due to the fact that C.3(sch) atoms have a higher probability

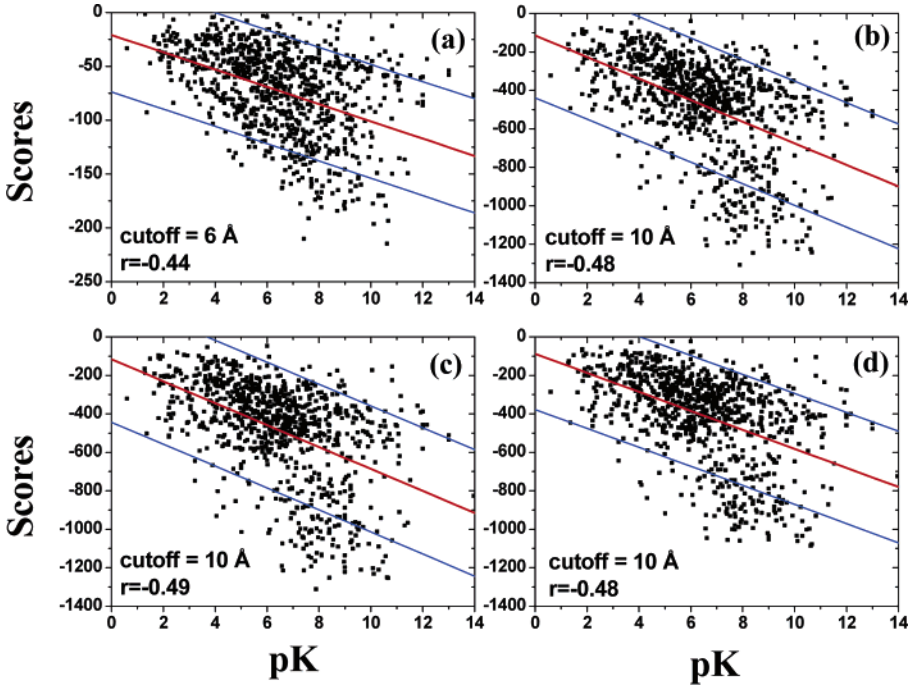

Figure 3. Correlation of the scores with experimental binding affinity data for the 896 complexes test set. Results of the scoring function based on new protein atom types and Mol2 atom types for the ligands are in (a) using cutoff = 6 Å, (b) using cutoff = 10 Å, (c) using cutoff = 10 Å and *B*-factor-augmented scoring function. Results of the scoring function based on new protein atom types and the atom types of PMF for the ligands using cutoff = 10 Å are in part d. Note that the scale of the *y*-axis in part (a) is different from those in (b), (c), and (d). In all, red lines are the correlation line fit and blue lines are 60% confidence lines.

directly interacting with ligands than the Cα or C*β* protein atoms, which are often shielded by protein side chain atoms. Hence, side chain and backbone atoms in proteins with the same bond order could have different knowledge-based potentials, suggesting the importance in making the distinction between them.

For the interaction between polar atoms, sharper peaks, such as those in the pair distributions of N.4-O.co2 and N.am-O.co2, are typically found (Figure 2). These sharper peaks correspond to deeper attractive potentials among these atom pairs. Interestingly, two potential wells, at around 2.8 Å and 5 Å, are found between the nitrogen atom and the O.co2. The well at shorter separation is deeper when the nitrogen (N.4) is charged, indicating that salt bridge interaction is generally stronger than a neutral hydrogen bond interaction. The potential between O.bk and N.4, however, shows only one potential well at a typical hydrogen bond distance (2.8Å).

Influences of Inclusion of the Atomic Mobility of Protein Atoms on Pairwise Distribution Functions and Potentials. We next analyzed the influences of inclusion of the atomic mobility of protein atoms on pairwise distribution functions and potentials.

As indicated in Methods, a Gaussian distribution based upon the isotropic *B*-factors available in crystal structures of protein−ligand complexes is used to describe the statistical probability of each protein atom instead of a fixed position. As can be seen, the pairwise distribution functions, including the atomic mobility of protein atoms (red lines in Figure 2), closely trace the path of those without inclusion of the atomic mobility of protein atoms (black lines in Figure 2), but in much smoother fashion. Of note, since the atomic *B*-factors used in our implementation contain information of local atomic environment, it is different from the running average or a triangular weighting as implemented in BLEEP[8] and Drugscore,[10] where each atom is treated equally.

In addition to the smoothing effect, the shapes of pairwise distribution functions for polar atom pairs show significant changes. For example, without inclusion of protein atomic mobility, sharp peaks are observed for the pairwise distribution functions among polar atom pairs at short distances and inclusion of protein atomic mobility leads to much smoother peaks in each case (Figure 2i−m).

Inclusion of protein atomic mobility also yields a less repulsive potential among the nonpolar interactions. For example, the repulsive part of the potential for distances less than 4 Å shifts to a closer separation in Figure 2e−h. The effects on polar interaction are more dramatic in three aspects: (1) the potentials become less repulsive and the potential wells are less deep; (2) the minimum of the potential shifts by about 0.2 Å to a closer separation; and (3) more pronounced repulsive potentials at a very close separation not seen in the single protein conformation model are now observed. For both nonpolar and polar interactions, the shapes of the potentials at long ranges (greater than 4 Å) are smoother but otherwise are generally not affected.

Interestingly, inclusion of the *B*-factor in our knowledge-based potentials resembles the strategies to soften the interaction potentials used by Ferrari et al.[22] and Stahl et al.[23] in their docking studies. They introduced empirical parameters to soften the interaction potentials to account for receptor flexibility when performing docking simulations. Although beyond the scope of the current study, it will be very interesting to test the effectiveness of our current scoring function in docking studies.

Evaluation of M-Score against 896 Protein−Ligand Complexes. We evaluated M-Score using 896 protein−ligand complexes, whose experimentally determined binding affinities are available from our PDBbind.[14,15]

For these 896 protein−ligand complexes, their high-resolution crystal structures have been determined.[15] Therefore, for the prediction of binding affinities for these 896 protein−ligand

Table 3. Performance of the Current Scoring Function against Different Protein Families in Correlating Scores and Experimentally Determined Binding Affinities. Only the protein family (wild type or mutants) with at Least Ten Protein−Ligand Complexes from the Test Set Are Included in This Analysis

		linear correlation coefficient (r)		
protein names	numbers in the data set	this work + Mol2	this work + Mol2 + *B*-factor	this work + Muegge
FK506 binding protein	10	−0.92	−0.92	−0.91
thermolysin	12	−0.83	−0.83	−0.83
trypsin	63	−0.76	−0.74	−0.71
urokinase-type plasminogen activator	14	−0.75	−0.78	−0.69
tyrosine phosphatase	27	−0.72	−0.71	−0.69
neuraminidase	12	−0.65	−0.70	−0.69
ribonuclease T1, A	13	−0.60	−0.62	−0.61
thrombin	37	−0.55	−0.55	−0.53
carbonic anhydrase II	39	−0.54	−0.54	−0.54
antibody Fab fragment	31	−0.52	−0.51	−0.52
protocatechuate 3,4-dioxygenase	10	−0.49	−0.41	−0.67
HIV-1, HIV-2 protease	79	−0.43	−0.43	−0.43
oligopeptide binding protein	28	−0.36	−0.38	−0.36
tyrosine kinase	12	−0.32	−0.34	−0.34
endothiapepsin	15	0.08	0.07	0.12
coagulation Factor X, Xa	15	0.20	0.18	0.33
streptavidin	21	0.48	0.46	0.40
average over 17 families of proteins	438	−0.452	−0.456	−0.451

complexes, the protein atom mobility information as quantified by the atomic *B*-factors could be included in the calculations of scores using the M-Score program. However, in a practical drug design project, such information is not available until the designed ligand is synthesized and a high-resolution crystal structure of protein−ligand complex is determined. For this reason, our current evaluation of M-Score used a single protein conformation based upon the protein−ligand complex coordinates available from the PDBbind database for these 896 protein−ligand complexes.

In Figure 3, we showed the correlation between the experimental binding affinity data of these 896 complexes and the scores calculated from M-Score based on our atom types for proteins and the Tripos Mol2 atom types for ligands.

Cutoff distances varying from 6 to 12 Å for the pair potential functions have been used previously by different groups.[7,8,10,11,13] Generally, there is a tradeoff between speed and performance with different cutoff distances. In our case, when the cutoff distance is increased from 6 Å to 10 Å, the linear correlation coefficient (r) improves slightly from −0.44 to −0.48 (cf. Figures 3a and 3b). Perhaps more significantly, a steeper slope for the correlation line is observed. This corresponds to a wider spread in scores for protein−ligand complexes and could make the scoring function more effective for differentiating protein−ligand complexes with similar binding affinities. These results suggest that the longer range interactions between 6 Å to 10 Å are important for some protein−ligand complexes in our diverse validation data set. However, increase of the cutoff distance from 10 Å to 12 Å had no significant effect on the correlation coefficient or the slope (data not shown).

Surprisingly, inclusion of protein atomic mobility in the scoring function only improved slightly the overall linear correlation coefficient (r) between the scores and experimental binding affinities for these 896 structurally diverse protein−ligand complexes (Figure 3b versus Figure 3c).

Our recent study[24] has shown that when a set of 800 protein−ligand complexes was employed to test the predictive power of 14 scoring functions, all of them produced poor to modest correlations between the calculated scores and experimental

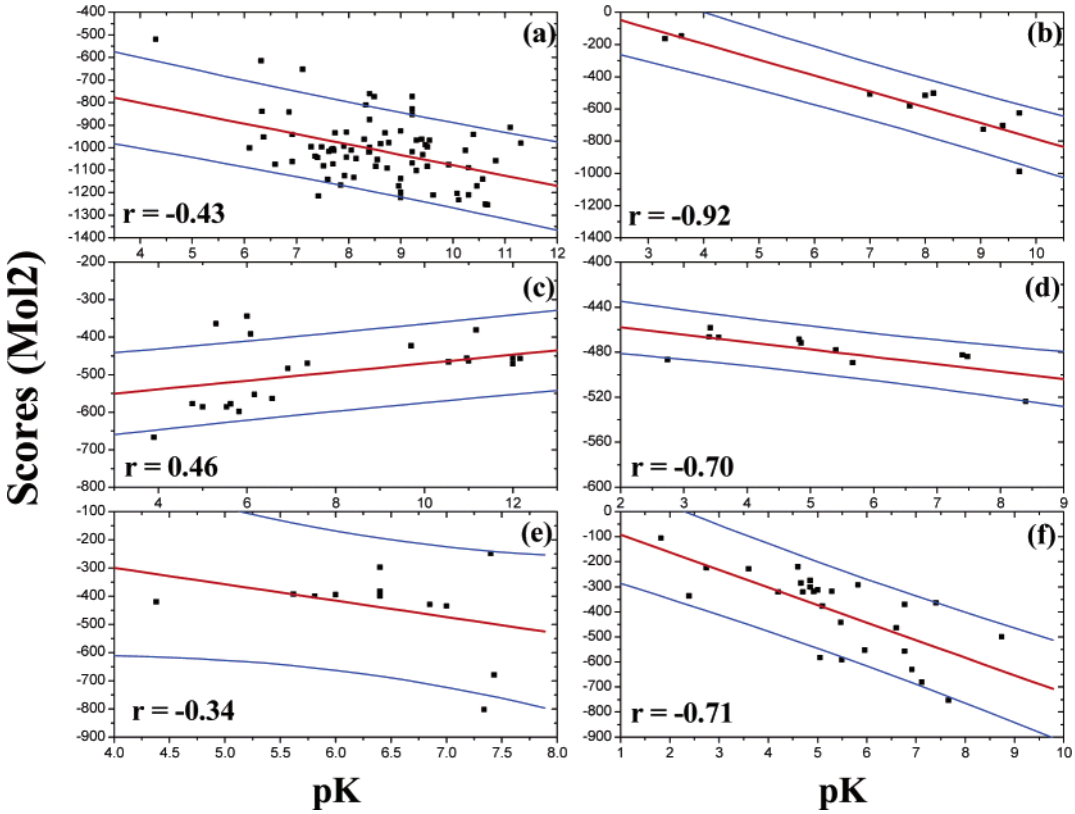

Figure 4. Examples: (a) HIV-1/-2 inhibitor (the largest single protein family set in current data set), (b) FK506−binding protein (the best case), (c) streptavidin (the worst case), (d) neuraminidase (slight improvement with *B*-factor augmentation), (e) tyrosine kinase (protein family phosphorylating specific tyrosine residues), and (f) tyrosine phosphatase (protein family removing phosphate groups from tyrosine residues). Results here are from the *B*-factor-augmented scoring function. The Tripos Mol2 atom type is used for ligands.

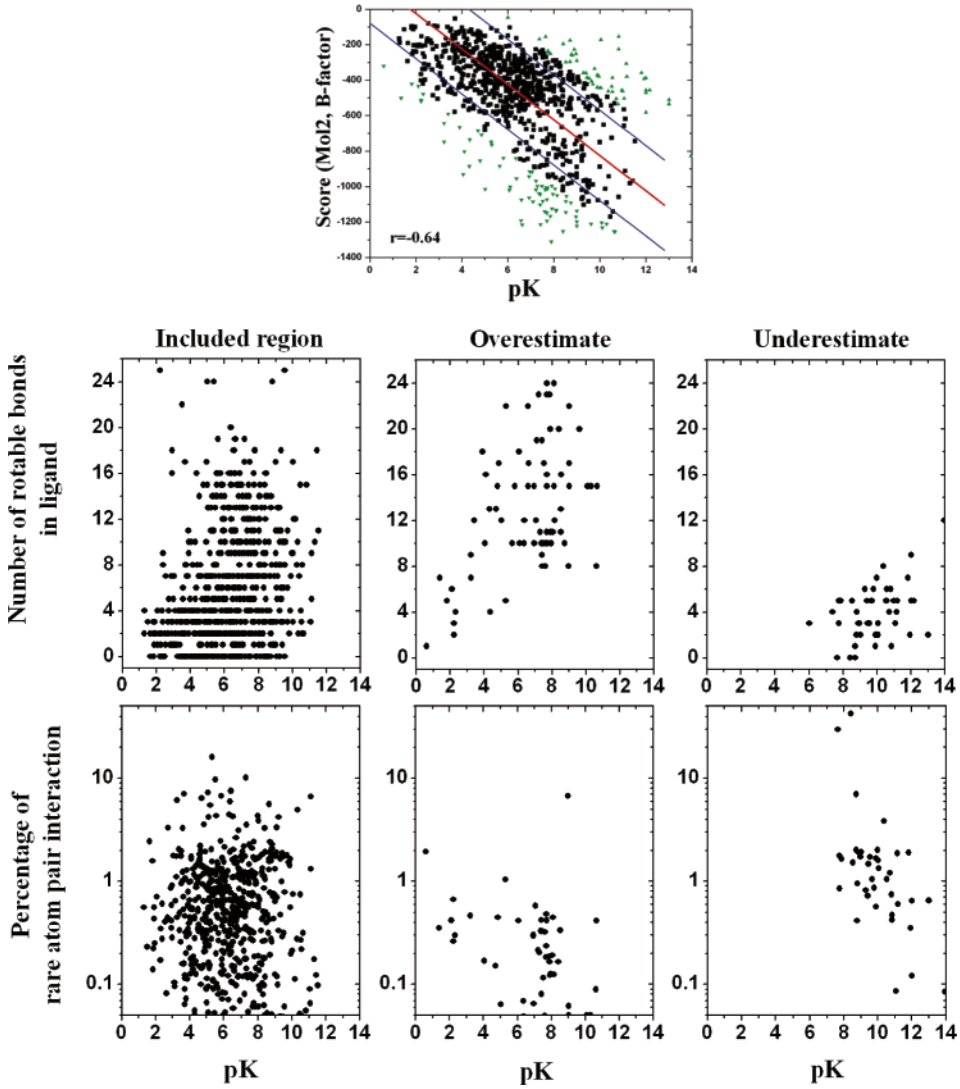

Figure 5. Outlier analysis: The analysis is based on results of M-Score. The rare atom pairs are those with less than 1000 distances in atom pairs in Table 3. Removing the under- (up-triangle) and overestimate (down-triangle) cases increases the linear correlation coefficient (r) from −0.49 to −0.64. 768 complexes are in the included region.

binding affinity data. It was found that the apparent disappointing results were primarily due to very large errors for a relatively small number (5%) of protein−ligand complexes.[24] We also found that M-Score was unable to predict the binding affinities for a small-percentage of protein−ligand complexes and produced very large errors for these outliers (Figure 3c).

When the atom types of PMF are used for ligands, the linear correlation coefficient (r) between the scores and experimental binding affinity is the same as that obtained when the Tripos Mol2 atom types are used (Figure 3d). Interestingly, comparison of the scatter plots between Figures 3b and 3d indicates that several overestimated cases with pK values from 6 to 11 become less problematic when the atom types of PMF are used for ligands.

Performance of the Scoring Function on 17 Protein Families. In structure-based lead discovery and optimization, one is often interested in the performance of a scoring function for its predictive power for the relative binding affinities of ligands bound to the same protein, or proteins in the same family. For this purpose, we evaluated M-Score against 17 protein families from the 896 protein−ligand complexes (Table 3). Each protein family consists of at least 10 different protein−ligand complexes. Proteins in each family may include the native and mutated forms.

We found that M-Score performs well for 6 of the 17 protein families with a linear correlation coefficient (r) varying between −0.70 and −0.92. These protein families include FK506 binding proteins, thermolysin, trypsin, tyrosine phosphatase, urokinase-type plasminogen activator, and neuraminidase. Our scoring function performs modestly well for 4 other protein families with a linear correlation coefficient (r) varying between −0.51 and −0.62. Among these 17 protein families, HIV-1/-2 proteases contain the largest number of protein−ligand complexes (79 entries) and the linear correlation coefficient (r) is −0.43, comparable to that for the overall data set. M-Score performs extremely poorly for coagulation factor X(Xa), endothiapepsin, and streptavidin with no or even anti-correlation.

With inclusion of the B-factor in the scoring function, the linear correlation coefficients (r) are the same or slightly improved in 9 out of the 17 protein families for two different atom typing schemes listed in Table 3. However, inclusion of the B-factor in the scoring function did not significantly improve those poorly predicted cases. Scoring functions using certain atom typing schemes may perform significantly better than others for specific protein families. For example, scoring functions using the atom types of PMF for ligands perform significantly better than those using the Mol2 atom typing scheme for ligands bound to protocatechuate 3,4-dioxygenase.

Interestingly, our current scoring function yields different performance for tyrosine kinase and tyrosine phosphatase. While good correlations are observed for tyrosine phosphatase complexes regardless which atom typing schemes are used (Table 3), poor correlations are found for tyrosine kinase complexes. Examples of the scatter plots for HIV-1/-2 proteases, FK506-binding protein, streptavidin, neuraminidase, tyrosine kinase, and tyrosine phosphatase are shown in Figure 4.

Analysis of the Under- and Overestimated Cases. To identify factors contributing to the extremely poorly performing cases in the 896 complexes, we defined the under- and overestimated cases with predicted errors greater or less than 2 standard deviation (2 σ). This led to the identification of 77 overestimated and 51 underestimated cases, representing 14% of the total complexes.

The 77 overestimated cases include 5 in MHC class I proteins, 39 in HIV-1/-2 proteases, and 7 in endothiapepsin. The 51 underestimated cases include 8 in streptavidin, 3 in thrombin proteins, 9 in coagulation factor X/Xa, and 8 in carbonic anhydrase II proteins. Not surprisingly, by excluding the under- and overestimated cases, the linear correlation coefficient (r) improves significantly from -0.48 to -0.64 for the remaining 768 complexes (Figure 5, top panel).

The number of rotable bonds for ligands is related to the conformational entropy penalty associated with complex formation. Since the number of rotable bonds for ligands is not included in M-Score, we evaluated if it plays a role for these under- and overestimated cases. Since knowledge-based potential scoring functions rely on statistics, we also evaluated if rare atom pairs between proteins and ligands contribute to these outliers.

The rotable bonds for ligands were calculated using X-Score.[6] Indeed, overestimated cases on average have a larger number of rotable bonds than underestimated cases or complexes with predicted errors less than 2 σ (Figure 5, middle panel). Thus, inclusion of ligand conformational entropy penalty incurred upon complex formation may improve the correlation for those overestimated cases.

One of the deficiencies in the knowledge-based potential scoring function is the insufficient number of certain atom pairs to generate their pairwise distribution functions and thereby their interaction potentials. Although 2331 protein−ligand complexes were used in the development of M-Score, 35% of the atom pair potentials were built by less than 1000 of the atom pairs, which may be considered as rare atom pairs. Although these rare atom pair interactions typically account for less than 10% of the total atom pair interactions for each complex (Figure 5, bottom panel), they may still be highly significant. Thus, using an even larger data set of protein−ligand complexes with diverse chemical structures of ligands to increase the number of these rare atom pairs may improve the performance of M-Score.

Summary

A new knowledge-based potential scoring function, which we named M-Score, was developed using 2331 crystal structures of protein−ligand complexes with resolution better than 2.5 Å. In M-Score, a Gaussian distribution based upon the isotropic *B*-factors available in crystal structures of protein−ligand complexes is used to describe the positional statistical probability for each protein atom instead of a fixed position. This leads to a more realistic representation of distance distribution for each atom pairs as compared to the use of a single, fixed distance when constructing knowledge-based potentials and a smoothing effect on the pairwise distribution functions. M-Score is validated using 896 complexes whose experimentally determined binding affinities were available and not included in the 2331 data set. The overall linear correlation coefficient (r) between the calculated scores and experimentally determined binding affinities (pK_i or pK_d) for the 896 validation data set is -0.49. Evaluation of M-Score against 17 protein families showed that good to excellent correlations were obtained for 6 protein families, the FK506 binding protein, neuraminidase, thermolysin, trypsin, tyrosine phosphatase, and urokinase-type plasminogen activator, with linear correlation coefficients (r) between -0.70 and -0.92. This indicates that M-Score can be used to predict the relative binding affinities of ligands to the same protein target for some protein families. However, there were poor correlations between the calculated scores and experimentally determined p*K* values for 7 protein families, including the coagulation factor X/Xa, endothiapepsin, and streptavidin. Analysis of those under- and overestimated cases suggested that neglect of the ligand conformational entropic changes upon binding, and insufficient statistics for rare atom pair interactions, may be some of the deficiencies in M-Score.

Acknowledgment. The authors would like to thank Dr. G. W. A. Milne for his critical reading of the manuscript.

Supporting Information Available: Comparisons of correlation for protein−ligand complexes shown in Figure 3 using different atom types for ligands and those without inclusion of atomic mobility of the proteins, numbers of atom pairs based on different atom types for protein and ligand atoms, the list of PDB entries of the 17 protein families, and the M-Score parameters. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Rev. Drug Discovery* **2004**, *3*, 935−949.
(2) Tame, J. R. H. Scoring Functions: A view from the bench. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 99−108.
(3) 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.* **2004**, *15*, 411−428.
(4) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(5) Böhm, 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.
(6) 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.
(7) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplied potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(8) 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.
(9) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, 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*, 1165−1176.
(10) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactgions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(11) Ishchenko, A. V.; Shakhnovich, E. I. Small Molecule Growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2000**, *45*, 2770−2780.
(12) DeWitte, R. S.; Shakhnovich, E. I. 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.

(13) Zhou, H.; Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Prot. Sci.* **2002**, *11*, 2714−2726.
(14) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(15) Wang, R. Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(16) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(17) Rodes, G. *Crystallography made crystal clear*, 2nd ed.; Academic Press: New York, 2000.
(18) Coppens, P. *X-ray charge densities and chemical bonding*; Oxford University Press: New York, 1997; pp 28−29.
(19) Tripos, Inc., St. Louis, MO 63144.
(20) Abramowitz, M.; Stegun, I. A., Eds. *Handbook of mathematical functions*; National Bureau of Standards; Dover: New York, 1974.
(21) Acton, F. S.; Watkins, W. *Numerical methods that work*; The Mathematical Association of America: Washington, DC, 1997.
(22) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *J. Med. Chem.* **2004**, *47*, 5076−5084.
(23) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(24) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein−ligand complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114−2125.

JM050043W