An Improved PMF Scoring Function for Universally Predicting the Interactions of a Ligand with Protein, DNA, and RNA

Xiaoyu Zhao,[†,#] Xiaofeng Liu,[‡,#] Yuanyuan Wang,[‡] Zhi Chen,[‡] Ling Kang,[†] Hailei Zhang,[†] Xiaomin Luo,[‡] Weiliang Zhu,[‡] Kaixian Chen,[‡] Honglin Li,*[,†,‡,§] Xicheng Wang,*[,†] and Hualiang Jiang*[,‡,§]

Department of Engineering Mechanics, State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology, Dalian 116023, China, Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zu Chong Zhi Road, Zhangjiang Hi-Tech Park, Shanghai 201203, China, and School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China

Received December 20, 2007

An improved potential mean force (PMF) scoring function, named KScore, has been developed by using 23 redefined ligand atom types and 17 protein atom types, as well as 28 newly introduced atom types for nucleic acids (DNA and RNA). Metal ions and water molecules embedded in the binding sites of receptors are considered explicitly by two newly defined atom types. The individual potential terms were devised on the basis of the high-resolution crystal and NMR structures of 2422 protein−ligand complexes, 300 DNA−ligand complexes, and 97 RNA−ligand complexes. The optimized atom pairwise distances and minima of the potentials overcome some of the disadvantages and ambiguities of current PMF potentials; thus, they more reasonably explain the atomic interaction between receptors and ligands. KScore was validated against five test sets of protein−ligand complexes and two sets of nucleic-acid−ligand complexes. The results showed acceptable correlations between KScore scores and experimentally determined binding affinities (log K_i's or binding free energies). In particular, KScore can be used to rank the binding of ligands with metalloproteins; the linear correlation coefficient (R) for the test set is 0.65. In addition to reasonably ranking protein−ligand interactions, KScore also yielded good results for scoring DNA/RNA−-ligand interactions; the linear correlation coefficients for DNA−ligand and RNA−ligand complexes are 0.68 and 0.81, respectively. Moreover, KScore can appropriately reproduce the experimental structures of ligand−receptor complexes. Thus, KScore is an appropriate scoring function for universally ranking the interactions of ligands with protein, DNA, and RNA.

1. INTRODUCTION

The past decade witnessed the successful application of emerging computational toolkits in the drug discovery and functional research of biological molecules. One of the focuses is to estimate the binding affinities for the processes of molecular recognition, including protein−ligand, nucleic-acid−ligand, protein−protein, and protein−nucleic-acid interactions.[1–3] However, due to the complexity of biological molecular recognition mediated by electrostatic interaction, van der Waals (VDW) interaction, hydrophobic interaction, hydrogen bonding (H-bond), and solvent effects, obtaining an accurate energy (scoring) function to precisely predict the binding affinity is still pending. The more complicated methods, such as free energy perturbation[4] and thermodynamic integration,[5] and their simplified version, the linear interaction energy approach,[6] are impractical in large-scale virtual screening or macromolecule docking simulation, owing to the unbearable computational expenses. Therefore, there is an urgent need to develop scoring functions that may predict binding affinity more effectively and accurately.

The current scoring functions for molecular docking can be classified in terms of their derivation approaches.[7] Force-field-based scoring functions apply classical molecular mechanics energy functions, approximating the binding affinity of receptor−ligand complexes as the summation of VDW and electrostatic interactions.[8–10] The solvent effect is usually taken into account using a distance-dependent dielectric function, and several solvent models based on continuum electrostatics have been developed.[11,12] Hydrophobic contributions are generally assumed to be proportional to the solvent-accessible surface area. A drawback of this approach is that the energy landscapes associated with force-field potentials are generally rugged, so that a minimization is required prior to the energy evaluation. The empirical methods are based on the widely accepted assumption that the contribution of binding affinity can be incorporated with uncorrelated terms.[13–16] The coefficient of each term is derived from multiple variable regression analysis methods by fitting available experimentally determined affinity values of a series of ligand−receptor complexes. Although such

* To whom correspondence should be addressed. Tel.: +86-21-50805873. Fax: +86-21-50807088. E-mail: hlli@mail.shcnc.ac.cn (H.L.), guixum@dlut.edu.cn (X.W.), or hljiang@mail.shcnc.ac.cn (H.J.). Please address correspondence to Prof. Hualiang Jiang, Ph.D., at Shanghai Institute of Materia Medica, Chinese Academy of Sciences.
† Dalian University of Technology.
‡ Chinese Academy of Sciences.
§ East China University of Science and Technology.
# These authors contributed equally to this work.

10.1021/ci7004719 CCC: $40.75
Published on Web 06/14/2008

kinds of methods are appealing for their ease in derivation, they are dubious for scoring new ligands structurally different from those used in the training set. The knowledge-based scoring methods that emerged in recent years are in essence designed to reproduce the experimental structures (binding poses) of ligands binding to receptors rather than the binding energies. The potential (more accurately referred to as potential mean force, PMF) of such a method is directly derived, according to the inverse Boltzmann law, from the statistical analysis of different types of atom pairs encoded in available crystal complex structures. These methods capture every interaction term implicitly (including solvation and the entropic effect) with an obvious advantage that it can be constructed without any knowledge of the binding data, and thereby they can be used to score novel ligands that are different from molecules in the training sets. A key issue of a knowledge-based potential is the reference state, which determines the weights among the various probability distributions. Several approaches to derive these potentials have been proposed with different definitions of the reference state, the receptor and ligand atom types, and the training sets.[17–23]

The success of statistical potential in predicting receptor−ligand binding affinity arouses the inspiration of extending this method to estimate protein−protein, protein−nucleic-acid, and nucleic-acid−ligand interactions. To this end, we have developed a more efficient and enhanced version of the PMF scoring function, which can evaluate not only protein−ligand but also DNA/RNA−ligand interactions. In this scoring function, the following improvements have been implemented: (1) the complexity of perceiving ligand atom types from Protein Data Bank (PDB) files has been circumvented by adopting the advantage of the unambiguous atom types defined in the mol2 format and adding newly defined atom types to characterize the metal−ligand and water−ligand interactions; (2) a special atom-typing scheme for nucleic acids (DNA and RNA) has been developed to characterize each individual base to capture the base-selectivity/-sensitivity of ligand binding. The atom pair interaction potentials were devised on the basis of several large-scale training sets. The scoring function based on the derived potentials, named KScore, was then validated against five test sets of protein−ligand complexes and two sets of nucleic-acid−ligand complexes for which the experimentally determined binding affinities (log K_i's or binding free energies) are available. The results showed acceptable correlations between KScore scores and experimental binding affinities. In particular, KScore can be used to rank the binding of ligands with metalloproteins; the linear correlation coefficient (R) for the test set is 0.65. In addition to reasonably ranking protein−ligand interactions, KScore also yielded good results for scoring DNA/RNA−ligand interactions; the linear correlation coefficients for DNA−ligand and RNA−ligand complexes are 0.68 and 0.81, respectively. This indicates that KScore is an appropriate scoring function to universally predict the interactions of ligands with protein, DNA, and RNA, elucidating that the redefined and supplementary atom types can extend the PMF scoring method to handle nucleic-acid−ligand interactions.

2. MATERIALS AND METHODS

2.1. The Potential Function. Similar to other knowledge-based scoring functions,[18–22] the potential function of KScore is defined by eq 1:

$$\mathrm{KScore} = \sum_{\substack{pl \\ r<r_{\text{cut-off}}}} A_{ij}(r) \tag{1}$$

where $A_{ij}(r)$ is the interaction term between a protein atom of type i and a ligand atom of type j, and r is the distance between atoms i and j. The sum is performed over all of the protein (or a nucleic-acid)−ligand atom pairs, pl, within a defined cutoff $r < r_{\text{cut-off}}$ in a database. $A_{ij}(r)$ is calculated in a similar way to PMF99,[22] which can be written as

$$A_{ij}(r) = -k_B T \ln\left[f^{j}_{\text{vol-corr}}(r)\,\frac{\rho^{ij}_{\text{seg}}(r)}{\rho^{ij}_{\text{bulk}}}\right] \tag{2}$$

where k_B is the Boltzmann constant, T is the absolute temperature, $f^{j}_{\text{vol-corr}}(r)$ is the ligand volume correction factor, $\Delta^{ij}_{\text{seg}}(r)$ is the number density of atom pair ij that occurs in a spherical shell with a thickness of Δr ranging from r to $r + \Delta r$, and $\Delta^{ij}_{\text{bulk}}$ expresses the number density when no interaction between i and j occurs. The ratio $\Delta^{ij}_{\text{seg}}(r)/\Delta^{ij}_{\text{bulk}}$ designates the radial distribution function of the atom pair ij. In detail, the individual term of radial distribution function can be evaluated by eq 3

$$\rho^{ij}_{\text{seg}}(r) = \rho^{ij}_{(r,r+\Delta r)}(r) = \sum_{pl} \frac{n^{ij}_{(r+\Delta r)}}{V_{(r+\Delta r)}(r)}, \quad \rho^{ij}_{\text{bulk}} = \sum_{pl} \frac{n^{ij}_{\text{bulk}}}{V(R)} \tag{3}$$

where $V_{(r+\Delta r)}(r)$ is the volume of the spherical shell with a thickness of Δr, $V(R)$ is the volume of the reference sphere with a radius of R, and $n^{ij}_{(r+\Delta r)}$ and n^{ij}_{bulk} are the numbers of occurrences of the protein−ligand atom pair ij in the spherical shell and reference sphere, respectively.

In principle, the interactions between protein/nucleic acid and ligand atoms are noncovalent so that the interaction does not frequently occur within covalent regions. Hence, ρ^{ij}_{bulk} could be invisible in these regions. A ligand volume correction factor similar to that of PMF99[22] has been proposed and applied to filter out the interactions of intraligands, and the desolvation and entropic terms can be treated implicitly. The ligand volume correction factor can be written as

$$\hat{f}^{j}_{\text{vol-corr}}(r) = \frac{\hat{\rho}^{ij}_{\text{bulk}}\,\rho^{ij}(r)}{\hat{\rho}^{ij}(r)\,\rho^{ij}_{\text{bulk}}} \tag{4}$$

where $\hat{f}^{j}_{\text{vol-corr}}(r)$ designates the available solvent/protein volume by ignoring the ligand−ligand interactions. The ligand volume-corrected number densities can be calculated using eq 5:

$$\hat{\rho}^{ij}(r) = \rho^{ij}(r)\,\frac{\rho^{kj}(r)}{\rho^{kj}(r) + \rho^{lj}(r)}, \quad \hat{\rho}^{ij}_{\text{bulk}} = \rho^{ij}_{\text{bulk}}\,\frac{\rho^{kj}_{\text{bulk}}}{\rho^{kj}_{\text{bulk}} + \rho^{lj}_{\text{bulk}}} \tag{5}$$

where $\rho^{kj}(r)$ designates the number density of protein atoms of type k around a ligand atom of type j in the spherical shells between r and $r + \Delta r$, and ρ^{kj}_{bulk} is the number density of the same atom pairs within an appropriate reference sphere with a radius of R. Thus, $\hat{f}^{j}_{\text{vol-corr}}(r)$ can be understood as the quotient of effective volumes taken by the ligand in the

Table 1. The 23 Ligand Atom Types Used in KScore

definition	notation	definition	notation
C.3	carbon sp^3	C.2	carbon sp^2
C.1	carbon sp	C.ar	cabon aromatic
C.cat	carbocation in a guadinium group	N.3	nitrogen sp^3
N.2	nitrogen sp^2	N.1	nitrogen sp
N.ar	nitrogen aromatic	N.am	nitrogen amide
N.pl3	nitrogen trigonal planar	N.4	nitrogen sp^3 positively charged
O.3	oxygen sp^3	O.2	oxygen sp^2
O.co2	oxygen in carboxylate and phosphate groups	S.3	sulfur sp^3
S.2	sulfur sp^2	S.O	sulfoxide sulfur
S.O2	sulfone sulfur	P.3	phosphorus sp^3
F	fluorine	Cl	chlorine
Br	bromine		

Table 2. The 28 DNA and RNA Atom Types

definition	notation	definition	notation
CnR	nitrogen connecting to sugar in cytosine ring	AND	nitrogen as hydrogen bond donor connecting to adenine ring
Can	nitrogen as hydrogen bond acceptor in cytosine ring	UnR	nitrogen connecting to sugar in uracil ring
CND	nitrogen as hydrogen bond donor connecting to cytosine ring	UnD	nitrogen as hydrogen bond donor in uracil ring
COA	oxygen as hydrogen bond acceptor connecting to cytosine ring	UOA	nitrogen as hydrogen bond acceptor connecting to uracil ring
TnR	nitrogen connecting to sugar in thymine ring	C2P	polar sp^2 carbon bonded to hetero atoms
TnD	nitrogen as hydrogen bond donor in thymine ring	C2F	nonpolar sp^2 carbon only bonded to carbon or hydrogen
TOA	oxygen as hydrogen bond acceptor connecting to thymine ring	C3F	nonpolar sp^3 carbon only bonded to carbon or hydrogen
GnR	nitrogen connecting to sugar in guanine ring	C3P	polar sp^3 carbon bonded to hetero atoms
GnA	nitrogen as hydrogen bond acceptor in guanine ring	O3	sp^3 oxygen in sugar
GnD	nitrogen as hydrogen bond donor in guanine ring	P	phosphorus
GND	nitrogen as hydrogen bond donor connecting to guanine ring	PO	phosphate oxygen
GOA	oxygen as hydrogen bond acceptor connecting to guanine ring	MET	metal ions
AnR	nitrogen connecting to sugar in adenine ring	OW	water oxygen
AnA	nitrogen as hydrogen bond acceptor in adenine ring	HH	hydrogen

reference sphere and in a certain spherical shell with a thickness of Δr. In the implementation of KScore, $\hat{f}^j_{\text{vol-corr}}(r)$ has been smoothed according to a set of spherical shells called segments, seg(r) (seg = 1, 2,..., $R/\Delta r$), consecutively separating the sphere with radius R into $R/\Delta r$ segments. Thus, the ligand volume correction factor can be practically calculated by eq 6:

$$f^j_{\text{vol-corr}}(r) = \begin{cases} \hat{f}^j_{\text{vol-corr}}(r) & \text{seg} \le m \text{ or seg} \ge R/\Delta r - m \\ \dfrac{1}{2m+1}\displaystyle\sum_{\text{seg}-m}^{\text{seg}+m} \hat{f}^j_{\text{vol-corr}}(r) & m < \text{seg} < R/\Delta r - m \end{cases} \quad (6)$$

where m is an adjustable variable and can be chosen according to the size of the training set or the requirement of the smoothing degree. As an example, the distance dependences of $f_{\text{vol-corr}}(r)$ values for five selected ligand atom types are shown in Figure S1 in the Supporting Information. In this study, we set $\Delta r = 0.2$ Å, $R = 12$ Å, and $m = 8$.

2.2. Atom Types. A total of 16 of the 17 protein atom types in KScore were from the PMF99/PMF04 scoring function.[17,22] Additionally, we introduced a metal atom type (MET) to specially characterize the metal ions buried in the protein binding sites (Table S1 in the Supporting Information). For ligands, we constructed 23 atom types based on the atom-type definitions of the mol2 format[24] (Table 1). These atom types are convenient for assignment by several existing programs such as Corina[25] and OpenBabel.[26]

On the basis of the definition of protein atom types in PMF04, we defined 28 nucleic acid atom types (Table 2). Each heteroatom embedded in the bases was denominated on the basis of its parent base name (A, T, C, G, or U), the base ring's structural characteristic, and its chemical property (hydrogen-bond acceptor or donor). The criteria for defining other atoms are similar to that of PMF04, reflecting the atom bond order and polarity states. Experimentally observed facts indicated that some ligands could recognize AT-rich or GC-rich sequences specifically.[27,28] Our scheme for defining the nucleic acid atom types was also designed to satisfy such special interactions. Met and OW atom types of protein were adopted directly to describe the metal ions and water molecules studded in the structures of nucleic acids.

2.3. Training Data Sets. The crystal structures of protein−ligand complexes were extracted from the PDB.[36] More attention has been paid to the high-resolution structures (<3.0 Å), especially those with experimental binding affinities which can be found in available protein−ligand binding databases such as LPDB,[29] PDBBind,[30,31] AffinDB,[32] PLD,[33] BindDB,[34] and PDTD.[35] Redundant records were eliminated unless the bound ligands were very diverse, and complexes of a protein bound with multiple ligands were treated as separate entries. As a result, a total of 2422 entries of protein−ligand complexes were collected for the training set (Table S2 in the Supporting Information). It should be noted that covalently bound ligands and peptide inhibitors remained, and their potentials were also included in KScore (see Results and Discussion).

To develop the scoring function for predicting nucleic-acid−ligand interaction, a special training set of nucleic-acid−ligand complexes was established. Different from DrugScore$^{\text{RNA}}$,[37] a scoring function for purely predicting ligand−RNA interactions developed on the basis of the data set of RNA−protein complexes, our training set only includes nucleic-acid−ligand (small molecule) complexes, of which

the structures were also extracted from the PDB. The metal ions embedded in the nucleic acid were treated as part of the binding sites. To collect as much information as possible to derive the potential function, we include NMR structures as well, and only the first structural model was adopted. Thus, 300 DNA−ligand and 97 RNA−ligand complexes were used to derive the distance-dependent pair potentials of KScore (Table S2 in the Supporting Information). Each ligand in the complexes was converted into the mol2 format.

2.4. Test Data Sets. We test the performance of KScore against three testing sets of protein−ligand complexes which have been tested by PMF04. In addition, two other high-resolution diverse sets of protein−ligand complexes isolated from the PDBBind were also selected as the testing sets. To further test the capability of KScore in predicting nucleic-acid−ligand interaction, nine DNA−ligand complexes and 15 RNA−ligand complexes with experimentally determined binding affinities were collected from the literatures. The PDB entries of these testing complexes are listed in Table S3 in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. The Potentials. The major motive to use mol2 atom types instead of PMF99/04 ones originates from the inherent incompletion and ambiguity of the atom-type scheme of PMF (Tables 1 and 2 and Table S1 in the Supporting Information). For example, PMF04 defines the imine nitrogen atom bonding to a hydrogen atom as either NA or ND, and the pyridine nitrogen atom can be assigned as either NP or NR. Moreover, some atom types such as sulfide sulfur have no appropriate definition in PMF99 (this flaw has been patched in PMF04). Another disadvantage of the existing PMF atom-typing scheme is the inconvenience and inaccuracy in the process of perceiving atom types from the PDB format. Accordingly, we adopted more universal and easier-to-convert mol2 typing rules to characterize the individual ligand atom types. Although there is no guarantee that all atom types can be assigned properly, the mol2 atom-typing scheme indeed disambiguates efficiently the atom-typing problems of PMF.

The potentials of representative atom-type pairs are shown in Figure 1. We compared the shapes of potential-distance curves of KScore with those of PMF04 by mapping the mol2 atom types to their counterparts in PMF04. The result indicates that our potentials are more reasonable than the PMF potentials. For example, the KScore potentials between positively charged atoms (e.g., N.4 and NC) and negatively charged atoms (e.g., OC and O.co2) are much deeper than the corresponding PMF potentials (Figure 1a and b). Similarly, the ND-O.2 and OA-N.am (or N.3) potentials exist in slightly deeper valleys at 2.8−3.2 Å (Figure 1c and d), properly describing the hydrogen-bond interactions between these atom types. Much deeper minima lie at ∼4.0 Å for the CF-C.3 and cF-C.ar potentials (Figure 1e and f), which commendably depict the hydrophobic interactions between these types of atoms. The PMF04 potentials almost cannot reflect the interactions of these atom types (Figure 1e and f), and the minima of the PMF99 potentials for these atom types are at ∼5.0 Å.[22] Thus, one has to artificially shift the carbon−carbon interaction to much smaller distance so as to retrieve a meaningful carbon−carbon potential gradient in the docking experiments by using PMF99 or PMF04 as the scoring function. The minima of KScore potential at 4.0 Å can reasonably eliminate the necessity of such artificial interference. In addition, KScore produced more appropriate potentials for the electrostatic repulsion interactions of OC-O.co2 (Figure 1g), in which the minimum shifts to 4−5 Å from 2.0 Å for the PMF04 potential. Another favored potential superior to PMF04 is the interaction between SA and C.3 atom pairs (Figure 1i). The prime potential minimum of PMF04 occurs at 2.2 Å, which accounts for forming covalent bonds between sulfur and carbon atoms. This is a common phenomenon in some proteases, especially those bearing cysteine as the catalytic center where the ligand may covalently bond to cysteine. However, the dominant interaction potential at shorter distances in PMF04 indicates that in most cases one has to shift the sulfur−carbon interaction to much larger distances artificially in order to eliminate the bias of covalent bonding atom pairs. KScore outperformed PMF04 in handling such cases. In accordance with the VDW interaction, KScore's SA-C.3 potential exhibits the prime potential minimum at ∼4 Å. Moreover, this potential also has the second minimum at 2.2 Å to describe the covalent bond for this atom pair.

To more explicitly take into account the solvent effect, we retained all of the water molecules in the original crystallographic complexes and treated them as parts of the protein. We also optimized the atom pair potentials between the water oxygen (OW) and ligand atoms. Figure 1j and k depict two of the representative potentials between OW and N.4 and between OW and O.co2, respectively. Again, these potential minima are considerably deeper than those of the OW−NC and OW−OC potentials of PMF04. Such results are clearly attributed to the fact that charged groups may get stabilized when exposed to the solvents due to the solvent effect.

Note that the PMF99[22] potential did not treat the metal ions as parts of the protein. However, recent research on metalloenzymes elucidated the crucial role of metal ions buried in the binding pockets or catalytic sites during the process of protein−ligand interaction.[38–40] Thus, several docking methods start to support automatic or manual assignment for the parameters of the metal ions.[41–43] Similar to the works by Wang et al.[18] and Zou and Huang,[20] we introduced a general metal ion type explicitly to characterize the particular interaction observed in metalloenzyme−ligand interactions. However, we only observed the notable interactions of metal ions of proteins with three carbon atom types (C.2, C.3, and C.ar) and two oxygen atom types (O.2 and O.3) of ligands. Thus, the potential interactions between metal ions and other ligand atom types may be discarded during the statistical process due to their infrequent occurrences. Figure 2a shows the potentials of MET interacting with O.2 and O.3 atoms. The valleys of both Met−O.2 and Met−O.3 potentials are located at ∼2.2 Å, which corresponds to considerable electrostatic interactions between positively charged metal ions and partial negatively charged oxygen atoms. This distance range also reveals the predominant chelated interactions between the lone-pair holding atoms and metal ions. Because the coordination bonds are much stronger than the nonbond interactions, Met−O.2 (O.3) interaction exhibits much deeper minima compared with the interaction between other atom pairs. On the other hand, the

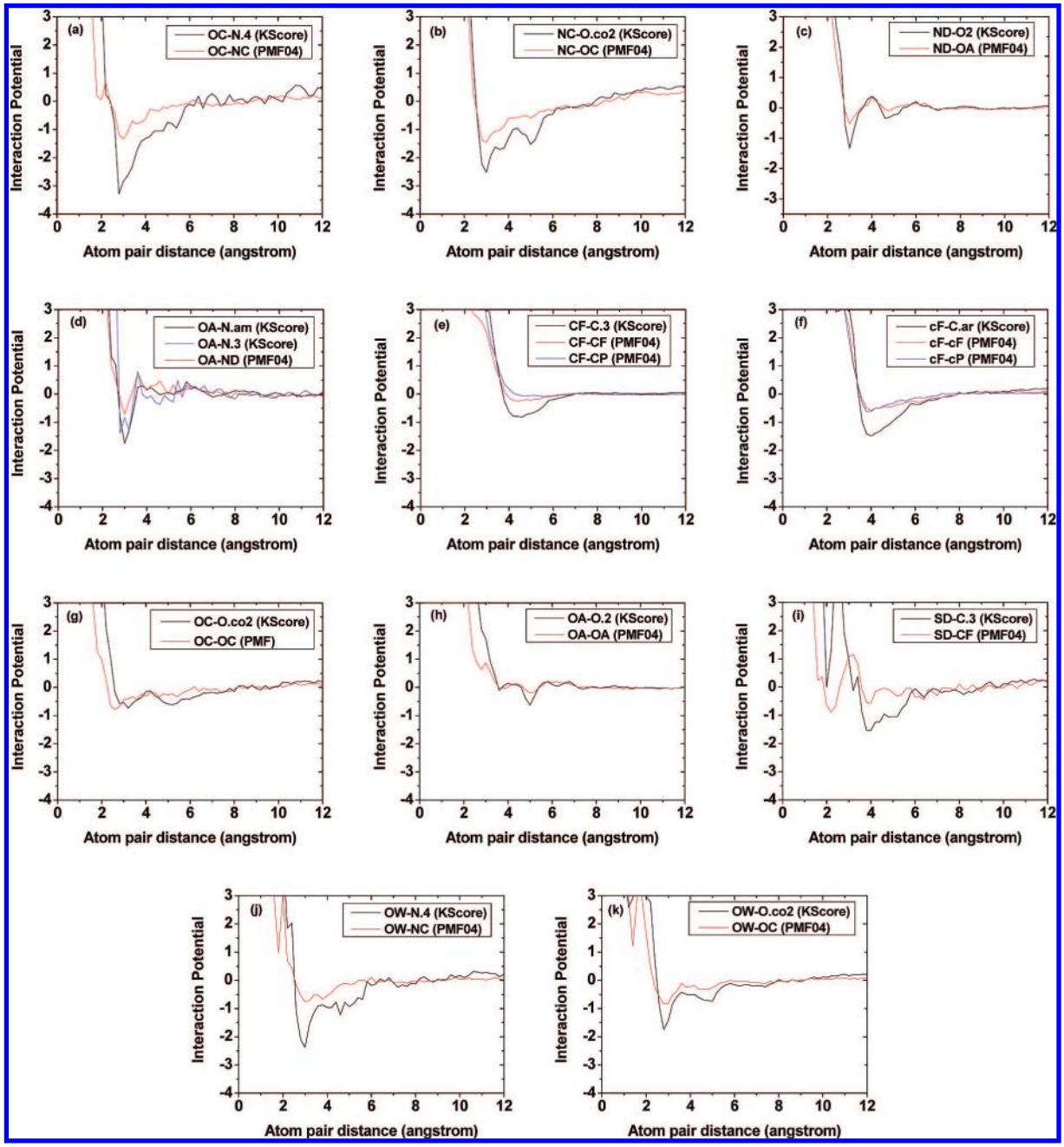

Figure 1. Comparison between KScore and PMF potentials for selected protein−ligand atom types. The four-letter (or five- and six-letter) code refers to the atom type pairs, where the first two letters before the dash refer to the protein atom types, and the rest of the letters indicate ligand atom types (see Table 1). The color of each curve is annotated explicitly in each figure, and all of the differences between KScore and PMF04 potentials are outlined in the text.

deeper potential minimum of the Met−C.ar interaction suggests that KScore may favorably describe the cation−π interaction, which is a potent, general noncovalent binding force observed in various biological systems.[44,45] Muegge also included a universal metal type to make the metal ions parts of proteins in recently released PMF04,[17] but confusingly some concrete metal atom types were also defined as ligands. To clarify this issue, we treated the metal ions as parts of either protein or DNA/RNA to emphasize the crucial roles that metal ions play in the binding pocket. Our strategy implies more biological and practical rationality in virtual screening targeting metalloproteins, because we explicitly considered the fact that ions in the binding pockets can anchor the compounds' binding poses.

Although a relatively larger data set embodying the complexes of DNA/RNA−ligand and DNA/RNA−protein/peptide may increase statistical frequencies of certain types of atom pairs, we only adopted the data of DNA/RNA−ligand complexes to develop the potentials because the major aim of this study was to devise a universal scoring function for predicting the interaction of ligands with both proteins and nucleic acids. The potentials of selected representative atom pairs of nucleic-acid−ligand interactions are shown in Figure 3. In general, the curve valleys locate at the same ranges as their counterparts observed in the protein−ligand interactions (Figure 1), 3.0−4.2 Å and 3.8−4.0 Å for PO−N.3/N.4 and C2F/C2P−C.ar potentials, respectively, indicating the reasonability of this set of potentials (Figure 3a and b).

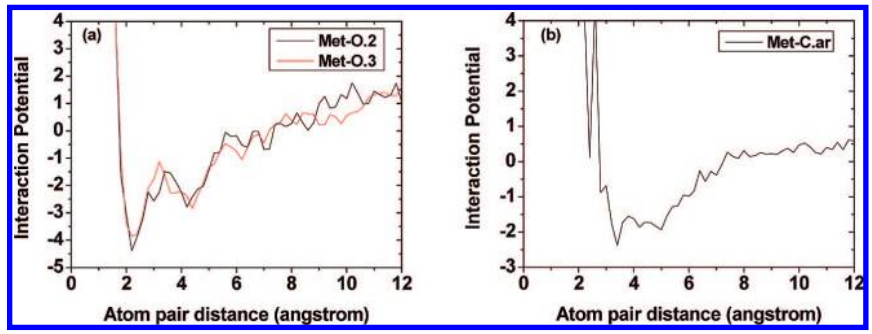

Figure 2. The KScore potentials between newly introduced metal ions containing atom types and three selected atom types of ligands. Met refers to the atom type of metal ions, and the rest of the letters refer to the ligand atom types (see Table 1 and Table S1 in the Supporting Information).

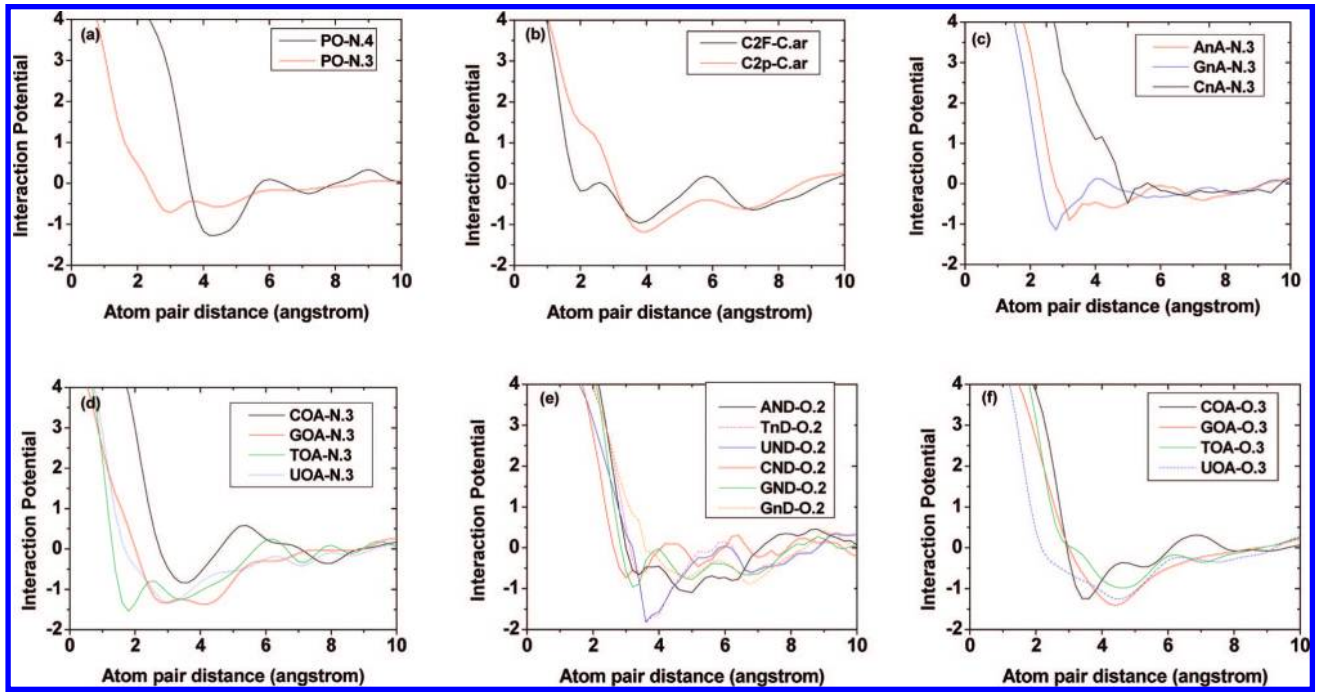

Figure 3. Selected KScore potentials for DNA/RNA−ligand atom type pairs. The five-letter or six-letter code refers to the atom type pairs, where the letters before the dashed line refer to DNA/RNA atom types, and the letters after the dashed line refer to the ligand atom types (see Tables 1 and 2). The color of each curve is annotated in every figure and is outlined in the text. To facilitate viewing, the curves presented here have been smoothed with the FFT filter method.

Moreover, the relatively deeper minima for the potentials of charged atoms and aromatic atoms indicate that these potentials may suitably describe the electrostatic and $\pi-\pi$ interactions, respectively.

Additionally, a new atom-type scheme was proposed to discriminate the specific atom pair interactions between different bases and ligands. Indeed, our potential can reflect the specificity of ligand−base interaction (Figure 3). As shown in Figure 3c, the potentials between the aromatic nitrogen atom types of three different bases as hydrogen-bond acceptors (AnA, GnA, and CnA) and putative the hydrogen-bond donor N.3 type have different minimum positions and depths. For GnA−N.3 and AnA−N.3 potentials, the smaller local minima positions (2.8 Å and 3.2 Å, respectively) and deeper wells correspond to their strong hydrogen bonding, and the larger minimum position of the CnA−N.3 potential (~5 Å) reveals a relatively weaker hydrogen bonding. This difference of potentials is associated with the inherent structural discrepancy of these three bases. For instance, the nitrogen atom at the N3 position of cytimidine (CnA atom type in Table 2) is usually not able to form hydrogen bonds with ligands due to its hydrogen bond with the partner N1 atom of guanine, while the N3 and N7 atoms (AnA and GnA atom types in Table 2) of guanine or adenine have the possibility to form hydrogen bonds with ligands (Figure S2 in the Supporting Information). The potentials of other atom pairs are shown in Figure 3d−f, which also reveal the interactions related to base specificity. This result indicates that KScore has considered the base specificity/sensitivity for recognizing ligands.

3.2. Predicting Protein−Ligand Complexes. We first applied the KScore to three test sets used previously by PMF99 and PMF04, then to a more diverse set containing 139 high-resolution structures of protein−ligand complexes derived from the PDBBind core subset[30,31] and an additional test set including 100 high-resolution structures used by Wang et al. for the performance to compare 11 existing scoring functions.[46] Tables 3 and 4 summarize the performance of KScore compared with other scoring methods; Figure 4 shows the respective correlations with the experimentally determined binding affinities.

Table 3. Correlation between Experimental Binding Affinity and KScore Prediction Against Five Protein Test Sets and Two Nucleic Acid Test Sets

				correlation coefficient (*R*)		
no.	test set[a]	figure	no. of complexes	KScore	PMF04[b]	PMF99[b]
1	Metalloenzyme	4a	15	0.65	0.78	0.76
2	Trypsin	4b	84 (47)[c]	0.72	0.87	0.73
3	HIV-1 protease	4c	58 (19)[c]	0.25	0.10	0.00
4	PDBBind core	4d	139	0.54	NA	NA
5	Wang's set	4e	100	0.49	NA	0.40[d]
6	DNA	5a	9	0.68	NA	NA
7	RNA	5b	15	0.81	NA	NA

[a] Set 1 was from the ChemScore data set of Eldridge et al.[15] Sets 2 and 3 were extracted from PDBBind.[30,31] Some complexes of these two sets have been used by PMF04.[17] Set 4 is a subset of PDBBind core, and only those having K_i values were tested. Set 5 was used by Wang et al. to evaluate the performance of 11 scoring functions.[46] Set 6 was derived from the test set used in PreDDICTA, and only those with experimental binding free energies were tested.[47] Set 7 is the test set used in developing DrugScore^{RNA} by Gohlke and Pfeffer.[37] [b] The correlation coefficients of PMF04 and PMF99 against individual test sets (if available) are listed for comparison (see text for details). NA indicates no information available for this set. [c] Trypsin and HIV-1 protease sets used in this study are larger in size than those used by PMF04 and PMF 99 (values in parentheses) due to a recent update. [d] The result of PMF99 against Wang's set is from ref 46.

Table 4. Correlations between Experimentally Determined Binding Affinities and Calculated Binding Scores of KScore and 15 Existing Scoring Functions against Wang's Test Set

scoring function	function type	correlation coefficient (*R*)	reference
ITScore	iterative score	0.65	20
X-Score	empirical	0.64	46
DFIRE	knowledge-based	0.63	20, 48
DrugScore^{CSD}	knowledge-based	0.62	19, 20
DrugScore^{PDB}	knowledge-based	0.60	19, 20
Cerius2/PLP	empirical	0.56	19, 20, 48
SYBYL/G-Score	force-field-based	0.56	20, 48
KScore	knowledge-based	0.49	this study
SYBYL/D-Score	force-field-based	0.48	19, 20, 46, 48
SYBYL/ChemScore	empirical	0.47	20,48
Cerius2/PMF	knowledge-based	0.40	20
DOCK/FF	force-field-based	0.40	20
Cerius2/LUDI	empirical	0.36	20
Cerius2/LigScore	force-field-based	0.35	20
SYBYL/F-Score	empirical	0.30	20
AutoDock	force-field-based	0.05	20

The first test set contains 15 metalloenzyme−ligand complexes,[14] KScore produced a correlation coefficient (*R*) of 0.65 (Figure 4a), which is slightly lower than those of PMF04 (0.78) and PMF99 (0.76) (Table 3). Although the roles metal ions played in KScore cannot be inferred from the result directly, the acceptable correlation result demonstrates that our strategy for treating the metal ions as parts of the proteins is reasonable. This result also illustrates that explicitly including metal ions may increase the capacity of the potential to discriminate the decoys that bear no or weak interactions with metal ions in virtual screening.

Sets 2 and 3 contain 84 trypsin protein−ligand and 58 HIV-1 protease−ligand complexes, respectively. The structures of these two sets are extracted from PDBBind database according to the classification criteria of PMF04.[17] The sizes of these two data sets are larger than those of PFMF04 because the entries of the 2007 version of PDBBind have been expanded compared with the old versions. For set 2, KScore gave an acceptable statistic result with a correlation coefficient (*R*) of 0.72 (Figure 4b), in comparison with PMF04 (0.87) and PMF99 (0.73) (Table 3). For set 3, the correlation of KScore improves slightly compared with PMF04; *R* increases to 0.25 from 0.1 for PMF04 (0.00 for PMF99)[17,22] (Figure 4c). However, the performance of KScore in this system is still poor, and more efforts should be made to improve the potentials.

Set 4 is the largest set used for KScore validation, and it contains 139 protein−ligand complexes with nonredundant sequences isolated from the PDBBind core subset. KScore revealed a correlation coefficient of 0.54 for this data set (Figure 4d), which indicates that KScore retains the capability of scoring ligands against diverse protein structures. The last set for protein−ligand complexes (set 5) includes 100 high-resolution structures used by Wang et al. in the benchmark comparison for 11 scoring functions, and the correlation of KScore to this data set is 0.49 (Figure 4e). To further test the scoring ability of KScore, we performed a comparison for KScore with 15 existing scoring functions against Wang's set. The result indicates that KScore performed better than average among the tested scoring functions in terms of correlation coefficient (Table 4).

However, knowledge-based scoring approaches intend to reproduce the experimental structures (binding poses) of ligands binding to receptors rather than the good correlation between the scoring data and experimental binding affinity. To test the ability of KScore to reproduce the experimental structures of ligand−protein complexes, we performed docking simulations on 69 ligand−protein complexes using our GAsDock program and the KScore scoring function. The result is listed in Table S4 in the Supporting Information. For the 69 ligand−protein complexes, KScore produced an average root-mean-square deviation (rmsd) of 2.27 Å, which

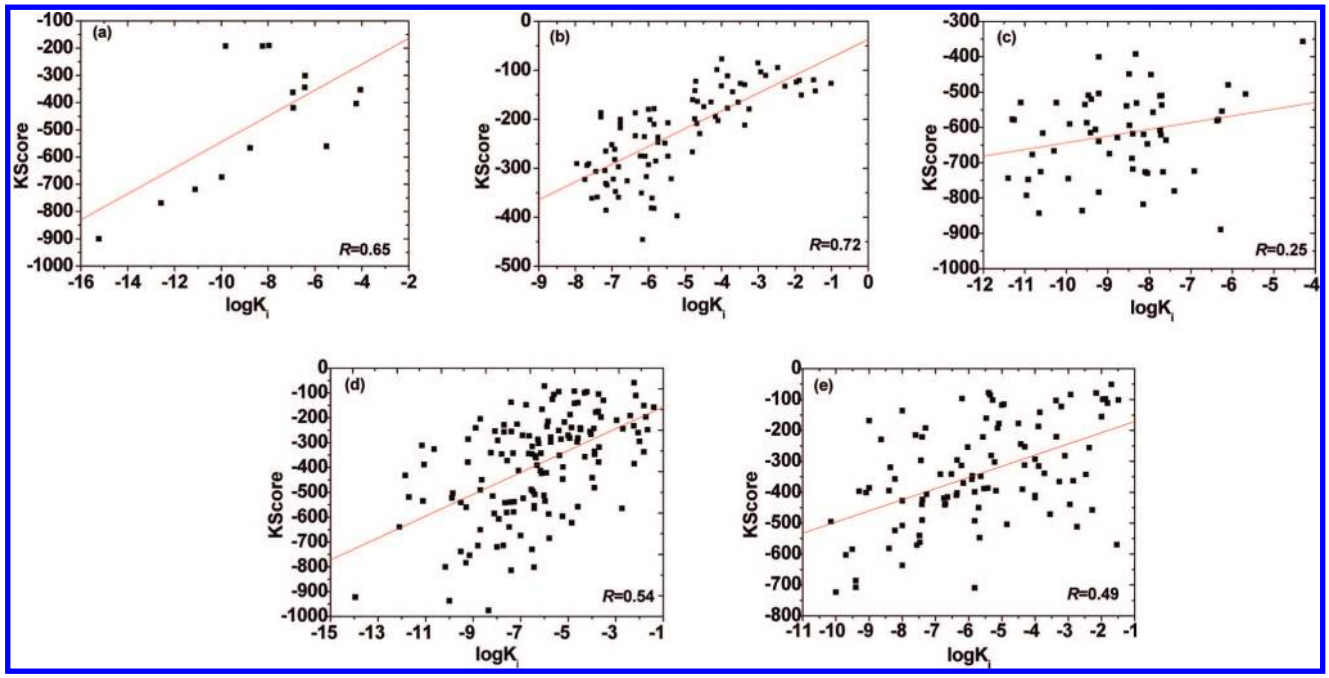

Figure 4. Correlation of KScore predicted values with experimentally determined binding affinities for various test sets, including (a) 15 metolloenzymes, (b) 84 trypsin proteins subset from PDBBind, (c) 58 HIV-1 proteases subset from PDBBind, (d) the PDBBind Core subset of 139 complexes (only those having K_i values have been selected), and (e) Wang's set (100 protein complexes used by Wang et al.; see Table 3).

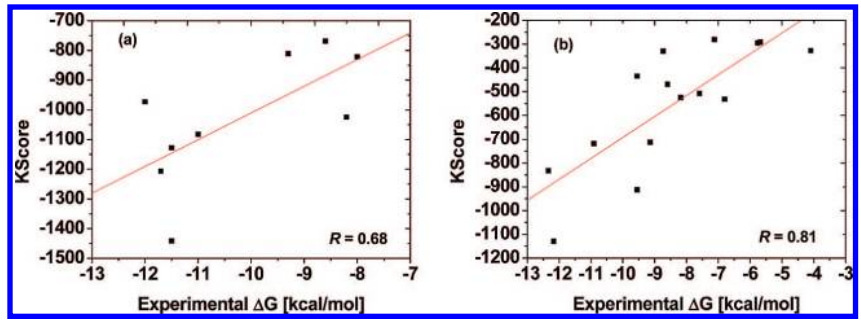

Figure 5. Correlation between experimentally determined binding free energies and KScore predicted values of (a) nine DNA−ligand complexes tested by PreDDICTA[47] and (b) 15 RNA−ligand complexes tested in DrugScore^{RNA},[37] see Table 3.

is lower than those of FlexX (2.66 Å) and DOCK (2.80 Å), and slightly higher than that of GOLD (2.14 Å). Moreover, KScore can reproduce 57% of the experimental structures with a rmsd of less than 2 Å, suggesting that the capability of KScore in reproducing experimental binding poses of ligands with the protein is as good as that of FlexX (57%) and better than that of DOCK (49%), but still weaker than that of GOLD (78%). This result indicates that KScore is an appropriate scoring function for docking simulation and virtual screening.

3.3. Predicting DNA/RNA−Ligand Complexes. We also estimated the capability of KScore to predict nucleic-acid−ligand binding by using nine DNA−ligand complexes and 15 RNA−ligand complexes as testing data sets (sets 6 and 7 in Table 3). These two data sets have been used to evaluate the performance of PreDDICTA[47] and DrugScore^{RNA}, respectively.[37] Figure 5 shows the calculated KScore values versus the experimentally determined binding free energies. For set 6 (DNA set), the correlation coefficient produced by KScore is 0.68 (Figure 5a), and that for 7 (RNA set) is 0.81 (Figure 5b). KScore gave better result than DrugScore^{RNA} in predicting ligand−RNA binding,[37] but the result in predicting ligand−DNA binding is not as good as that of PreDDICTA.[47] This result has not demonstrated which scoring function is advanced over others because the sizes of the two test sets are not large enough. Nevertheless, KScore is an acceptable scoring function in ranking both ligand−DNA and ligand−RNA bindings in comparison with existing scoring functions, specifically for DNA and RNA.

4. CONCLUSIONS

In this study, an improved knowledge-based scoring function, KScore, has been developed on the basis of several diverse training sets and a newly defined atom-type scheme. The large, diverse training sets as well as newly defined atom types contribute to the extraction of statistically significant atom pairwise potentials of Kscore. In comparison with the existing PMF potentials like PMF99 and PMF04, our pairwise potentials for different atom types have been significantly improved (Figures 13). In particular, our individual potentials better reflect the experimental phenomena such as the interaction distances and strengths of hydrogen bonding, electrostatic interaction, VDW interaction, cation−π interaction, and aromatic stacking. The redefined ligand atom types eliminated the ambiguity and complexity

inherent in the process of perceiving the ligand types from the PDB format. Tests of KScore on binding affinity prediction yield reasonable correlations with the experimental data for seven test data sets consisting of 396 protein−ligand complexes, nine DNA−ligand complexes, and 15 RNA−ligand complexes. Accordingly, KScore can be used to rank the interactions of protein−ligand, DNA−ligand, and RNA−ligand complexes. The devised potential terms for the atoms of water and metal molecules buried in the binding sites of proteins and nucleic acids (Figures 1k and 2) enhance the ranking capability of Kscore for metalloproteins and nucleic acids (Figures 4a and 5). Moreover, the involvement of the atom types of water molecules as parts of proteins or nucleic acids allows KScore to consider the solvent effect explicitly. Remarkably, the newly introduced atom-typing scheme for nucleic acids enables our potential not only to evaluate DNA/RNA−ligand interaction but also to recognize the base specificity for binding ligands (Figures 3e and f). This is one possible reason that Kscore produced better or acceptable results versus PreDDICTA[47] and DrugScore$^{RNA\ 37}$ in predicting DNA−ligand and RNA−ligand interactions, respectively (Figure 5).

The aim of developing a PMF scoring function is not to predict the absolute binding free energy for ligand−protein/RNA/DNA interactions, but to rank molecular recognition according to the binding pose and tendency.[22] To this end, KScore is an appropriate tool for distinguishing the strong binders from a series of compounds and can be applied to large-scale virtual screening as a promising scoring function (Table S4 in the Supporting Information). In addition, by appropriately modifying the atom-type scheme and diverse training sets, further improvements can be easily envisioned to extend KScore to the application of evaluating protein−protein and protein−DNA/RNA interactions. Recently, KScore has been integrated into our molecular docking program, GAsDock,[10] and the validation and application of KScore in docking simulation and virtual screening will be described in detail elsewhere.

ACKNOWLEDGMENT

The authors thank Prof. Ingo Muegge for providing the potentials of PMF04. This work was supported by the Special Fund for Major State Basic Research Project (Grants 2002CB512802 and 2004CB518900), the National Natural Science Foundation of China (Grants 10572033 and 30672539), the Shanghai Committee of Science and Technology (Grant 07dz22004), and the 863 Hi-Tech Program of China (Grant 2007AA02Z304 and Grant 2006AA02Z336). H.L. was also supported by Knowledge Innovation Program of the Chinese Academy of Sciences (Grant SIMM0709QN-09).

Supporting Information Available: The list of the 17 atom types of proteins is shown in Table S1; the PDB entries in the training sets for deriving the KScore function are listed in Table S2; the PDB entries included in the testing sets are listed in Table S3, and comparisons of KScore+GAsDock with FlexX, GOLD, and DOCK4 in reproducing experimental structures of ligands binding to proteins are listed in Table S4. Figure S1 shows the volume correction $f^{j}_{vol\_corr}(r)$ of five selected ligand atom types as a function of atom pair distances, and Figure S2 schematically represents the hydrogen-bonding interactions between atoms of five nucleic acid bases and ligand atoms. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **1997**, *72*, 1047–1069.
(2) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and "hot spots" for protein-ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115–144.
(3) McCammon, J. A. Theory of biomolecular recognition. *Curr. Opin. Struct. Biol.* **1998**, *8*, 245–249.
(4) Beveridge, D. L.; DiCapua, F. M. Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annu. Rev. Biophys. Biophys. Chem.* **1989**, *18*, 431–492.
(5) 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., III. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
(6) Hansson, T.; Marelius, J.; Åqvist, J. Ligand binding affinity prediction by linear interaction energy methods. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 27–35.
(7) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
(8) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195–202.
(9) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem.* **1992**, *13*, 505–524.
(10) Li, H.; Li, C.; Gui, C.; Luo, X.; Chen, K.; Shen, J.; Wang, X.; Jiang, H. GAsDock: a new approach for rapid flexible docking based on an improved multi-population genetic algorithm. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 4671–4676.
(11) Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4–16.
(12) Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.
(13) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein ligand complex of known 3-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.
(14) Böhm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.
(15) 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.
(16) Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427–440.
(17) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–902.
(18) Yang, C. Y.; Wang, R. X.; Wang, S. M. M-score: A knowledge-based potential scoring function accounting for protein atom mobility. *J. Med. Chem.* **2006**, *49*, 5903–5911.
(19) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScore^{CSD}-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction. *J. Med. Chem.* **2005**, *48*, 6296–6303.
(20) Huang, S. Y.; Zou, X. Q. An iterative knowledge-based scoring function to predict protein-ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, *27*, 1876–1882.
(21) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein-ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770–2780.
(22) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.
(23) 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.
(24) Clark, M.; Cram, R. D. Validation of the general purpose Tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(25) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
(26) Open Babel. http://openbabel.sourceforge.net/ (accessed Nov 11, 2007).
(27) Goodsell, D. S. Sequence recognition of DNA by lexitropsins. *Curr. Med. Chem.* **2001**, *8* (5), 509–516.
(28) Wemmer, D. E. Ligands recognizing the minor groove of DNA: Development and applications. *Biopolymers* **1999**, *52*, 197–211.
(29) Roche, O.; Kiyama, R.; Brooks, C. L., III. Ligand-protein database: linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592–3598.
(30) 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.
(31) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(32) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein-ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–526.
(33) Puvanendrampillai, D.; Mitchell, J. B. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **2003**, *19*, 1856–1857.
(34) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.
(35) Gao, Z.; Li, H.; Zhang, H.; Liu, X.; Kang, L.; Luo, X.; Zhu, W.; Chen, K.; Wang, X.; Jiang, H. PDTD: a web-accessible protein database for drug target identification. *BMC Bioinf.* **2008**, *9*, 104.
(36) 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.
(37) Pfeffer, P.; Gohlke, H. DrugScore^{RNA}-knowledge-based scoring function to predict RNA-ligand interactions. *J. Chem. Inf. Model.* **2007**, *47*, 1868–1876.
(38) Wasser, I. M.; De Vries, S.; Moënne-Loccoz, P.; Schröder, I.; Karlin, K. D. Nitric oxide in biological denitrification: Fe/Cu metalloenzyme and metal complex NO_x redox chemistry. *Chem. Rev.* **2002**, *102*, 1201–1234.
(39) Hightower, K. E.; Fierke, C. A. Zinc-catalyzed sulfur alkylation: insights from protein farnesyltransferase. *Curr. Opin. Chem. Biol.* **1999**, *3*, 176–181.
(40) Klabunde, T.; Krebs, B. The dimetal center in purple acid phosphatases. In *Metal Sites in Proteins and Models*, 1st ed.; Hill, H. A. O., Sadler, P. J., Thomson, A. J., Eds.; Springer: Heidelberg, Germany, 1997; Vol. 89, pp 177−198.
(41) Jain, T.; Jayaram, B. Computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. *Proteins* **2007**, *67*, 1167–1178.
(42) Irwin, J. J.; Raushel, F. M.; Shoichet, B. K. Virtual screening against metalloenzymes for inhibitors and substrates. *Biochemistry* **2005**, *44*, 12316–12328.
(43) Esposito, E. X.; Baran, K.; Kelly, K.; Madura, J. D. Docking substrates to metalloenzymes. *Mol. Simul.* **2000**, *24*, 293–306.
(44) Zacharias, N.; Dougherty, D. A. Cation-π interactions in ligand recognition and catalysis. *Trends Pharmacol. Sci.* **2002**, *23*, 281–287.
(45) Dougherty, D. A. Cation-π interactions in chemistry and biology: A new view of benzene, Phe, Tyr, and Trp. *Science* **1996**, *271*, 163–168.
(46) Wang, R. X.; Lu, Y. P.; Wang, S. M. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(47) Shaikh, S. A.; Jayaram, B. A swift all-atom energy-based computational protocol to predict DNA-ligand binding affinity and ΔT_m. *J. Med. Chem.* **2007**, *50*, 2240–2244.
(48) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A knowledge-based energy function for Protein-Ligand, Protein-Protein, and Protein-DNA complexes. *J. Med. Chem.* **2005**, *48*, 2325–2335.

CI7004719