JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

Knowledge-Based Scoring Functions in Drug Design: 3. A Two-Dimensional Knowledge-Based Hydrogen-Bonding Potential for the Prediction of Protein–Ligand Interactions

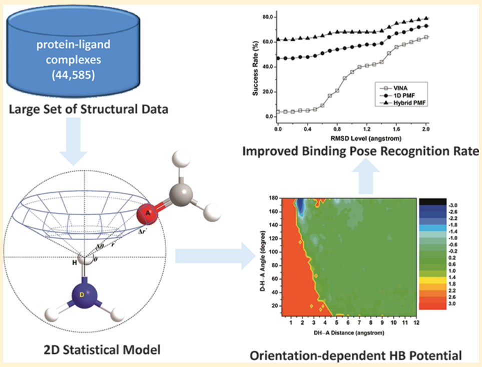
Mingyue Zheng,*[†,¶] Bing Xiong,[‡,¶] Cheng Luo,[†] Shanshan Li,[†] Xian Liu,[†] Qianchen Shen,[†] Jing Li,[†] Weiliang Zhu,[†] Xiaomin Luo,[†] and Hualiang Jiang*[,†]

[†]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[‡]State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

S *Supporting Information*

ABSTRACT: Hydrogen bonding is a key contributor to the molecular recognition between ligands and their host molecules in biological systems. Here we develop a novel orientation-dependent hydrogen bonding potential based on the geometric characteristics of hydrogen bonds observed in 44,585 protein–ligand complexes. We find a close correspondence between the empirical knowledge and the energy landscape inferred from the distribution of HBs. A scoring function based on the resultant hydrogen-bonding potentials discriminates native protein–ligand structures from incorrectly docked decoys with remarkable predictive power.



INTRODUCTION

Biomolecular recognition between ligands and their host molecules, typically proteins, serves as the physical basis of biological regulations. In the regulations, intermolecular hydrogen bonds (HBs) make a significant contribution to the stability of ligand-protein complexes due to their outstanding strength and ubiquitous nature. More importantly, the exquisite specificity of biological processes requires that the intermolecular interactions involved in the underlying recognition events are also specific.[1] Hydrogen bonding is by far the most important specific interactions,[2] and this is what makes it so important in the whole domain of biomolecular recognition. In the field of drug development, HBs are instrumental not only in mediating drug-receptor binding, but they also affect pharmacokinetic profile of a molecule.[3] Therefore, many computational efforts have been devoted to the energetic description of HBs.[1,4−13]

A hydrogen bond is formed when a positively polarized hydrogen atom bound to an electronegative donor atom penetrating the van der Waals sphere of an acceptor atom to interact with its lone pair electrons. This mechanism decides that the hydrogen bond has covalent, electrostatic, and van der Waals characters at the same time and spans an energy range from 40 kcal/mol to ~0.25 kcal/mol.[14] Another compelling aspect of the hydrogen bond is its angular preference. As frequently observed, the geometric arrangement of hydrogen bonds typically follows an orientation of the hydrogen toward the lone electron pairs of the acceptor atom. However, the location of the lone pair cannot be simply assumed based on the hybridization of the acceptor, as the hybridization state of the acceptor atom itself is, in turn, affected by hydrogen bond formation. Accordingly, such diversity in its energy components, directionality, and environment dependency sometimes causes computational modeling of hydrogen bonding energy landscapes a challenging problem.

Current approaches are generally classified into three groups: quantum mechanical (QM) calculations on reduced model systems,[15,16] molecular mechanics (MM) approaches,[17−19] and knowledge-based potentials derived from small molecule structure databases[6] or the Protein Data Bank (PDB).[4,5,7−9,12] For application to ligand-protein systems, different approaches have respective strengths and shortcomings. QM calculations can give the most accurate results, but they need to include electron correlation effects which are computationally expensive and can thus be applied to small model systems only. Empirical MM force fields are not restricted to the size of electronic structure and have been implemented in some scoring functions for ligand-protein interaction.[13] However, these approaches have the apparent shortcoming of limited transferability, resulting from parametrizations

Received: August 24, 2011
Published: October 15, 2011

dx.doi.org/10.1021/ci2003939 | *J. Chem. Inf. Model.* 2011, 51, 2994–3004

on specific systems. Inference of interaction energy landscapes directly from ligand-protein complexes has the advantage that it can model any behavior observed in experimentally determined structures, even if there is not a good physical understanding of the behavior. Compared with QM or force-field potentials, these knowledge-based potentials offer good compromise between the accuracy, general applicability, and the computational speed and tend to allow better handling of the uncertainties and deficiencies of computed interaction geometries.

The most common approach for developing knowledge-based potentials is to extract structural information from protein−ligand complexes and employ the Boltzmann law to transform the atom pair preferences into distance-dependent pairwise potentials. This type of potentials has been extensively explored during the last few decades and has received increasing interest with the rapid growing volume of high-quality structural data in the public domain.[20−28] However, the interaction model based on distance dependence only cannot fully represent the composite characters, especially the directionality, of a hydrogen bond.

To address this issue, Liu et al. derived distance- and angle-dependent potentials for various types of HBs and compared with QM calculations.[13] Some orientation hydrogen-bonding potentials were developed, comprising both distance-dependent energy term and angular-dependent energy components at the same time.[7,8,12] For these approaches, each energy term represents an individual knowledge-based potential, and different energy terms are considered independent of each other. (In this sense, these models are still one-dimensional in describing the HB interaction.) A few multidimensional models were also reported to investigate the dependence of HB energy on different combinations of distance and angular variables.[6,9] In probing such geometric preferences, multidimensional models require that the studied interactions are abundant enough to guarantee statistical significance, due to a more finely divided sampling space. Under the restriction, these multidimensional models are mainly designed for a certain type of HB form, not suitable for a general application. Besides, a large portion of current HB potentials are based on hydrogen-bonding geometries observed in structures of small organic molecules, proteins and protein−protein complexes, which cannot be readily applied to investigate HBs of protein−ligand binding interface.

Here we attempt to design a multidimensional statistical model to study the geometric and energetic preference of HBs. The Muegge's potential of mean force (PMF) analysis is performed on a large collection of protein−ligand complex structures, to derive a novel two-dimentional (2D) potential for various HB types. The resulted HB potentials are evaluated using two tests related to the prediction protein−ligand complexes.

■ METHODS

1. Preparation of Protein−Ligand Complex Structures. Initially, a local copy of the PDB database (release date: 20-Dec-2010) containing 69,940 entries was created via the RCSB FTP service. A qualified PDB entry for protein−ligand complex identification and extraction needs meet the following criteria: (1) it is determined by X-ray diffraction; (2) overall resolution is better than or equal to 3.0 Å; (3) it does not contain DNA, RNA, or multiple models; (4) it is composed by at least one protein molecule and one valid small-molecule ligand. For a PDB entry including one protein and multiple ligands, each individual ligand was grouped with the protein to compose a discrete complex, and separate pairs of coordinate files were saved. Consequently, one PDB entry may generate multiple protein−ligand complexes, while identical ligands in duplicate chains were only considered once (only the first one is kept). A molecule is considered as a valid protein if it is comprised of more than eight standard amino acids, and it becomes a valid ligand when it fulfills the following conditions: (1) it is not a part of cofactor, coenzyme, solvent, or buffer; (2) it should not contain any metal atoms or other uncommon elements; (3) it is an oligopeptide with less than eight residues; (4) it should not be covalently bound to any protein or other ligand atoms. Practically, both "ATOM" and "HETATM" records of the coordinate section were browsed to build molecules described in the PDB file, of which the structural components like cofactor or solvent were identified using predefined knowledge-based lists of PDB HET groups. Those various structural components were then appended to the structure file of the protein molecule. Finally, a cleanup step was applied to remove the complexes with severe steric clashes, which was defined as the distance between a protein atom and a ligand atom is shorter than 1 Å. All of the above extraction procedures were performed by a set of in-house python scripts.

The extracted structure file is not suitable for immediate use because typically it only consists of heavy atoms; it is necessary to specify the explicit positions of hydrogen atoms for either protein or ligand structures. The command line utility Prepwizard (The Protein Preparation Wizard) along with Schrödinger 2010 was used for the protein treatment. Specifically, missing atoms/residues were added; all waters were retained; all hydrogens in the original file were removed and readded; the exhaustive mode of the Protassign module was used to optimize the hydrogen-bonding network, rotate hydrogens (including waters), generate appropriate protonation and tautomerization states of His, and perform "chi-flips" in Asn, Gln, and His residues. In the case of hydroxyl and thiol-hydrogens, a set of hydrogen-assignment possibilities is selected based on the local electrostatic environment and then scored to determine the existence and quality of the hydrogen-bonding networks. For ligand, the OpenBabel (version 2.3) was used to adjust the ionization state and add hydrogens under neutral pH. In the end, the structure files of prepared protein were saved in the PDB format and ligand in SDF.

2. Derivation of 2D Hydrogen-Bonding Potential. Derivation of a potential of mean forces starts from the observation of a correlation between certain internal variables within a set of structures. The correlation is then converted into a PMF via an inverted Boltzmann formula

$$A(\Omega) = -k_B T \ln g(\Omega) \tag{1}$$

where Ω denotes a set of generalized degrees of freedom, $g(\Omega)$ is the probability density function over Ω that describes the correlation, k_B is the Boltzmann constant, and T is the absolute temperature.

Various models of interactions were explored during the developments of statistical potentials.[29] The most widely used ones are the distance-dependent potentials that account for interactions through radial distribution of protein−ligand atom pairs.[20−28] This type of potential is a typical one-dimensional (1D) model in that there is only one relevant degree of freedom that needs to determine the relative arrangement of two atoms, where $\Omega = \{r\}$. Take Muegge's method for example,[26,27] the atom pair potential

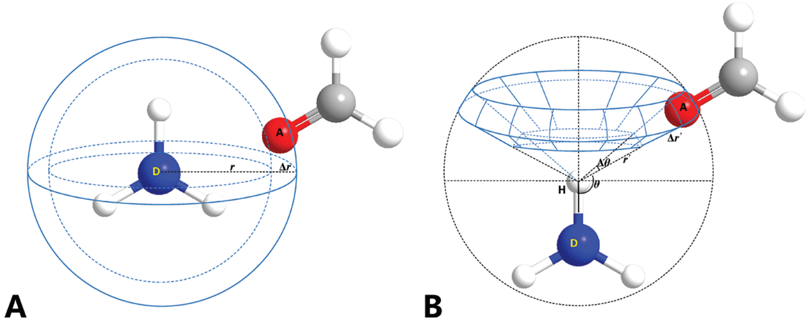


Figure 1. Definition of the coordinate system and the geometrical variables in the energy functions of (**A**) the 1D (distance dependent) and (**B**) 2D (the distance and angle dependent) HB potential models.

$A_{ij}(r)$ at distance r is obtained by

$$A_{ij}(r) = -k_B T \ln g_{ij}(r) \tag{2}$$

where i and j denote the atom types of protein and ligand, respectively; $g_{ij}(r)$ represents the pair distribution function with form

$$g_{ij}(r) = \frac{\rho_{ij}(r)}{\rho_{ij,\,bulk}} \tag{3}$$

where $\rho_{ij}(r)$ is the number density of a protein—ligand atom pair of type ij at a certain atom pair distance r; and $\rho_{ij,bulk}$ is the number density in a reference sphere with a radius of R (R = 12 Å). In practice, a correction factor $f_j\,(r)$ was introduced to the above function to account for the volume taken by ligand atom, so the effective pair distribution function writes

$$g^*_{ij}(r) = f_j(r)\frac{\rho_{ij}(r)}{\rho_{ij,\,bulk}} \tag{4}$$

and the corrected atom pair potential writes

$$A^*_{ij}(r) = -k_B T \ln g^*_{ij}(r) = A_{ij}(r) - k_B T \ln f_j(r) \tag{5}$$

Details about the calculation of $f_j(r)$ can be found elsewhere.[25,27] Figure 1A shows the representation of hydrogen bond interaction in the distance-dependent potential model. To calculate $\rho_{ij}(r)$, the spherical sampling space centered at the donor atom is divided into multiple layers. The thickness of each layer (Δr) is set to 0.1 Å.

In this study, we follow a similar formalism to derive the HB potentials and explore a different model of interactions. Given the atomic frame depicted in Figure 1B, two internal variables are introduced to describe the relative arrangement of the HB donor (D), acceptor (A), and the associated hydrogen (H): one is the H···A distance (r'), and the other is the D—H···A angle (θ). Clearly, compared to the model depicted in Figure 1A, this scheme brings an additional degree of freedom into the interpretation of the correlations they exhibit, i.e. $\Omega = \{r',\theta\}$. Therefore, for the two-dimensional (2D) model, one equation to calculate the atom pair potential at distance r' and angle θ is written as

$$A_{ij}(r',\theta) = -k_B T \ln g_{ij}(r',\theta) \tag{6}$$

where the subscription ij denotes the type of atom pair, of which the first letter represents the atom type of donor and the second acceptor. Since i and j, as defined in eq 2, have been used to discriminate between protein and ligand atoms (which usually have different atom typing schemes), respectively, eq 6 only applies to the hydrogen bonds formed by donor atoms of protein and acceptor atoms of ligand. For the remaining HBs formed by donor atoms of ligand and acceptor atoms of protein, the equation to calculate HB potentials is

$$A_{ij}(r',\theta) = -k_B T \ln g_{ij}(r',\theta) \tag{7}$$

For brevity, only the calculation of $A_{ij}(r',\theta)$ is described below, $A_{ji}(r',\theta)$ can be derived likewise. In eq 6, $g_{ij}(r',\theta)$ is the pair distribution function with form

$$g_{ij}(r',\theta) = \frac{\rho_{ij}(r',\theta)}{\rho_{ij,\,bulk'}} \tag{8}$$

where $\rho_{ij}(r',\theta)$ is the number density of observing donor—acceptor pair of type ij at a particular distance r' and angle θ, regardless if a hydrogen bond between them is possible; $\rho_{ij,bulk'}$ is the number density in a reference sphere with the H atom as the center, and the radius of the sphere is also set to 12 Å (R' = 12 Å). Figure 1B depicts the representation of hydrogen bond interactions in the 2D model. To calculate $\rho_{ij}(r',\theta)$, the sampling space centered at the hydrogen atom is divided into multiple spherical sectors and then divided in radial directions. The equation for computing each volume element $V(r',\theta)$ (the area enclosed by blue lines in Figure 1B), is given below

$$V(r',\theta) = \frac{4}{3}\pi[(r' + \Delta r')^3 - r'^3]\sin\left(\theta + \frac{\Delta\theta}{2}\right)\sin\left(\frac{\Delta\theta}{2}\right) \tag{9}$$

where the bin width $\Delta r'$ and $\Delta\theta$ were set as 0.2 Å and 5°, respectively. The probability density function $\rho_{ij}(r',\theta)$ is defined as

$$\rho_{ij}(r',\theta) = \sum \frac{n_{ij}(r',\theta)}{V(r',\theta)} \tag{10}$$

where $n_{ij}(r',\theta)$ is the number of interactions within the volume element; $\rho_{ij,bulk'}$ is defined as

$$\rho_{ij,\,bulk'} = \sum \frac{N_{ij}}{V(R')} \tag{11}$$

where N_{ij} is the count of interactions of ij type observed throughout the entire sampling space of volume $V(R')$

$$V(R') = 4\pi R'^3/3 \tag{12}$$

Based on the definitions provided in PMF04[26] and our previous implementation,[23] a set of 17 protein and 34 ligand atom types were used for both the 1D potential and 2D HB potential development. For the 2D HB potential, only the PMF04 atom types that can form hydrogen bonds were considered. A full list of these atoms types was summarized in the table in the Supporting Information.

3. Application of 2D HB Potentials in Scoring Function. The derived 1D potentials can be easily used for developing a scoring function for prediction of protein–ligand interactions, in which the PMF score (x) of a protein–ligand complex is calculated as the sum over all protein–ligand atom pair potentials $A_{ij}^{*}(r)$ at distance r

$$x = \sum_{\substack{kl \\ r < r^{ij}_{cut-off}}} A^{*}_{ij}(r) = \sum_{\substack{kl \\ r < r^{ij}_{cut-off}}} [A_{ij}(r) - k_B T \ln f_j(r)] \tag{13}$$

where kl is a protein–ligand atom pair of type ij, and $r^{ij}_{cut-off}$ is the distance at which atom pair interactions are truncated. The cutoff was set to 6 Å for carbon–carbon atom type interactions, and 9 Å for other interactions. In contrast, the derived 2D HB potentials are not ready for direct scoring function development because they only depict the hydrogen bonding interactions. In this study, a hybrid scoring function was designed to incorporate both 1D and 2D HB potentials. The main idea of this approach is straightforward and computationally easy to implement: under the assumption of additivity of atom-pair interaction energies, all the hydrogen bonding interactions are described by the 2D model and others by 1D. Since the same atom typing scheme was followed by both the 1D and 2D interaction models, this strategy can be conveniently applied as follows: Given a protein atom k with type i and a ligand atom l with type j, there are six relevant criteria that decide if the pairwise contribution of kl to the hybrid PMF score should come from the 1D or 2D model, and which equation should be used for the potential calculation: (I) k and l are either N, O, or S atoms; (II) k has at least one hydrogen atom m attached; (III) l has at least one hydrogen atom m' attached; (IV) the distance between m and l is shorter than 12 Å; (V) the distance between m' and k is shorter than 12 Å; (VI) the distance between k and l is shorter than 9 Å. With the above-described conditions, the unified equation to calculate the hybrid PMF score x_h is written as

$$x_h = \begin{cases} \sum_{kl} [A_{ij}(r', \theta) - k_B T \ln f_j(r)], \text{if I\&II\&IV\&VI;} \\ \sum_{kl} [A_{ji}(r', \theta) - k_B T \ln f_j(r)], \text{if I\&III\&V\&VI;} \\ \sum_{kl} [A_{ij}(r', \theta) + A_{ji}(r', \theta) - k_B T \ln f_j(r)], \text{if I\&II\&III\&IV\&V\&VI;} \\ \sum_{\substack{kl \\ r < r^{ij}_{cut-off}}} [A_{ij}(r) - k_B T \ln f_j(r)], \text{otherwise.} \end{cases} \tag{14}$$

This set of equations aims to address intermolecular interactions between the protein and the ligand. For a given ligand atom the interaction energy to all other ligand atoms is not considered. Therefore, the term $-k_B T \ln f_j(r)$ introduced in the 1D PMF potential (eq 5) is followed here, to correct the reference state and filter out intraligand interactions. Another point of note is that some atoms can be both an HB acceptor and donor, sometimes at the same time, e.g. OD, ND, and OW atoms. Figure 2 provides a schematic representation of two potential hydrogen bonds formed between a pair of atoms with type ij. This situation corresponds to that all the above-mentioned six conditions are met simultaneously. Thus, according to eq 14, the hybrid PMF between the atom pair is calculated as follows: $A_{ij}(r_1', \theta_1) + A_{ji}(r_2', \theta_2) - k_B T \ln f_j(r)$. Here, the contributions of these HB interactions are considered independent to each other and hence additive; the ligand volume correction is only made once because the atom pair only involves one heavy atom of ligand.

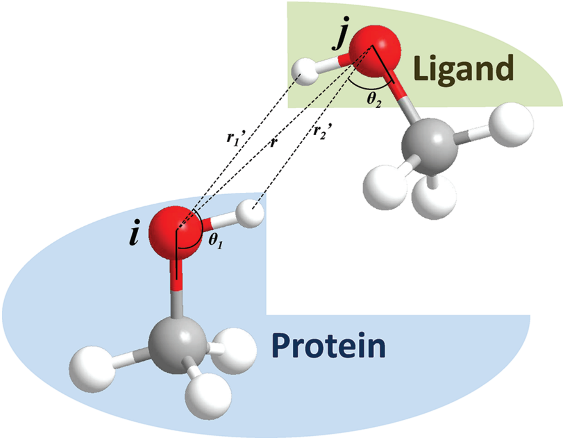


Figure 2. Schematic representation of two potential hydrogen bonds simultaneously formed between a pair of atoms with type ij.

Both the performance on binding affinity prediction and binding mode identification were evaluated to test the success of the hybrid PMF score. Two test sets were collected: One is the "core set" of the PDBBind (version 2009) database containing 219 complexes,[30,31] and the other is Wang's set[32] containing 100 complexes. These two sets are relatively large and contain diverse protein–ligand complexes. To test the performance on binding affinity prediction, the widely used Spearman's rank order correlation coefficient (R_s) was calculated, with detailed definitions provided elsewhere.[23] For comparison, the coefficient was obtained for PMF99, PMF04, and the 1D PMF derived in this study. To test the performance on binding mode identification, the following procedures were performed: for each ligand–protein complex in the test sets, up to 20 discrete putative ligand binding poses were first generated using the docking program AutoDock Vina (version 1.1.2), with the maximum number of binding modes to produce (*num_modes*) set to 20, the maximum energy difference between the best binding mode and the worst one (*energy_range*) to 10 kcal/mol, and the exhaustiveness level of the global search to 10. These decoy poses plus the native one in cocrystallized structure, making a binding pose ensemble for each complex in the test sets, were separately assessed by their hybrid PMF scores. Next, for each scoring function the highest ranked modes and their root-mean-square deviation (*rmsd*) values from the respective crystal structures were collected for the whole test set. For comparison, the 1D PMF and the score of AutoDock Vina were also tested as controls. The cumulative success rates of different scores were plotted to show the distribution of the best predicted poses with respect to the *rmsd* values.

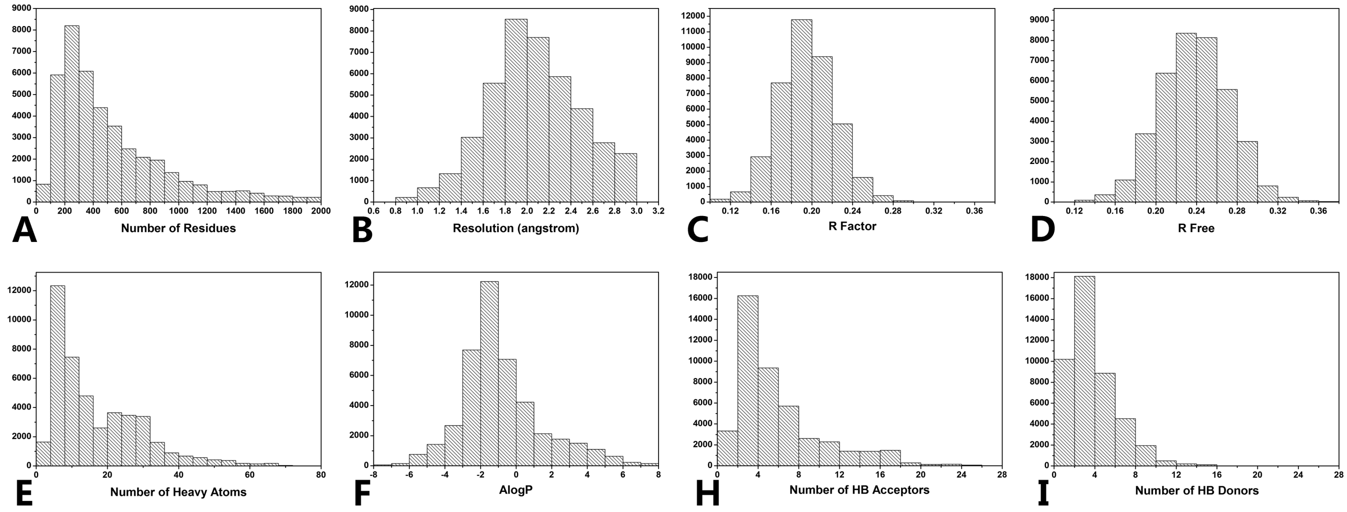


Figure 3. Histograms describing the statistics of the structural database: (A) protein primary sequence length; (B) crystallographic resolution; (C) working set R factor; (D) free R factor; (E) number of heavy atoms of ligand; (F) AlogP; (H) number of HB acceptors of ligand; (I) number of HB donors of ligand.

■ RESULTS AND DISCUSSION

1. Database of Protein−Ligand Complex Structures. The final database comprises 44,585 protein−ligand pairs encompassing 29,300 PDB entries. Several aspects of the database statistics are illustrated in Figure 3. Most of the protein data result from structures of 100−400 residues solved at resolutions between 1.8 and 2.2 Å with the most typical R and free R factors being 0.20 and 0.24, respectively. The drug-likeness of ligand was not the primary consideration in preparing the data set. Compared to the original 1D statistical model, the algorithm behind the current 2D model requires significantly increased data (which will be discussed in the next section). Therefore, one guiding idea of the data set preparation is to collect as many structures, especially the structures involving HB interactions, as possible. Following the procedures described in the Methods section, the resulted database of ligands has a large portion of small, lipophilic molecules. As shown in Figure 3, the number of heavy atoms of ligands has a widespread distribution with its maximum amplitude located at 5 and a notable enrichment at 20−30. The most frequent AlogP values of ligands fall in the range of −3.0 to 1.0, and the most frequent number ranges of hydrogen bond donors and acceptors are in [0,6] and [2,8], respectively.

2. Properties of Donor−Acceptor Pair Distribution. The statistical significance of knowledge-based potentials increases with the amount of data evaluated. Generally, for a 1D PMF potential, about 500 to 1000 occurrences per atom pair type were considered necessary to yield sufficiently smooth potential curves, depending on the size of sampling space and the number of bins.[20,27] It means that there need to be about 10 occurrences on average in each bin, to provide a significant potential of mean force. In our 2D interaction model, the spherical sampling space is divided in both axial and radial directions, which result in up to 2160 (12 Å/0.2 Å × 180°/5°) bins. Therefore, we conclude that the sufficient frequency for the evaluation of the HB interaction energies is about 20,000 for each atom pair. If the total number of occurrences of a specific donor−acceptor pair type in all sampling volume elements was smaller than 20,000 (lg $N_{ij(ji)}$ <4.3), we set $A_{ij(ji)}(r',\theta) = 0$ kcal/mol in all volume elements for the pair type. That is, in the 2D model we ignored a particular pair type if it had statistically insufficient data. The contributions of such null pair types to the hybrid PMF score were obtained from the 1D interaction model, in which the sufficient frequency threshold is set to a much lower value of 1000. A full list of the qualified donor−acceptor types and their occurrences can be found in the Supporting Information.

Table 1. Logarithm of Selected HB Donor−Acceptor Pair Occurrences in the Databases of Protein−Ligand Complexes Used To Derive the 2D PMF

		protein HB donor atom types (i)		
	ligand HB acceptor atom types (j)	NC	ND	OD
$\log_{10}N_{ij}$	OC	5.45	5.99	5.21
	OA	5.32	5.94	5.15
	OD	6.00	6.55	5.83
	NA	4.14	4.80	3.98
	ND	5.26	5.88	5.06
		ligand HB donor atom types (j)		
	protein HB acceptor atom types (i)	NC	ND	OD
$\log_{10}N_{ij}$	OC	5.22	5.22	5.68
	OA	5.89	5.96	6.37
	OD	5.19	5.20	5.72
	NA	4.72	4.76	5.19
	ND	5.92	5.98	6.39

Table 1 shows the statistics of a selected set of typical HB donor−acceptor atom pairs in this database. The number of atom pair occurrences varies among the atom types. The largest number of atom pairs occurred for the interaction between the protein HB donor type ND and the ligand HB acceptor type OD (lg N_{ij} = 6.55). Generally, typical hydrogen-bond interactions have enough observations to derive reliable 2D potentials, except

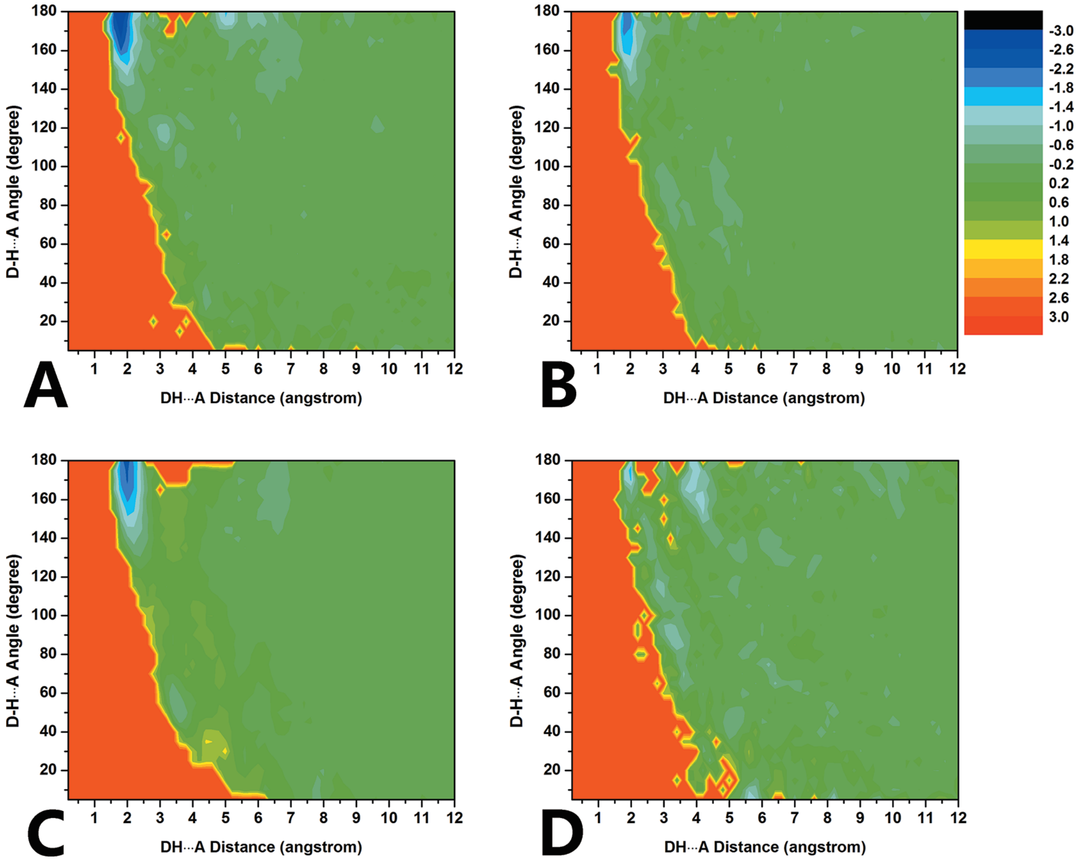


Figure 4. Heat maps of 2D HB PMF energy for a few donor−acceptor atom pairs involving hydrogen bonding interactions of different strength: (**A**) OD_lOA_p, (**B**) OD_lNA_p, (**C**) ND_lOA_p, and (**D**) ND_lNA_p. The six-letter code refers to the atom pair types, where the first two capitalized letters refer to the donor atom type and the last two refer to the acceptor atom type; the subscripts p and l indicate the protein and ligand atom type, respectively.

for the ligand acceptor type NA, of which two pair types, NCNA (lg N_{ij} = 4.13) and NDNA (lg N_{ij} = 3.98), fail the criterion of lg $N_{ij(ji)}$ >4.3. A full list of HB donor−acceptor atom pairs in the database and their respective observation numbers was summarized in the table of Supporting Information. Of those 96 pair types in all, only 34 (35.4%) have insufficient data for the 2D HB model, and three (3.1%) for the 1D model. Such coverage suggests that most combinations of the intermolecular HB donor and acceptor atom types are adequately sampled in the current knowledge base.

3. Characteristics of 2D HB Potentials. The $A_{ij(ji)}(r',\theta)$ energy heat maps for a few selected atom pair types are shown in Figure 4, in which larger negative energy values were represented by cold colors and positive values by warm colors. Figure 4A depicts the interactions between OD (oxygen as HB donor) and OA (oxygen as HB acceptor). If we set the threshold energy for hydrogen bonding to −1.4 kcal/mol, corresponding to the light blue color of the map, we may find that the hydrogen bond interaction between the pair is restricted to a small blue region defined by r' between 1.6 and 2.4 Å and θ in the range of 150−180°. No bonds are formed for θ smaller than 150°. These results clearly demonstrate the angular specificity in forming hydrogen bonds and reflect the superiority of the current 2D model in characterizing the HB interactions. Obviously, the traditional distance-dependent potentials are not capable of judging the existence of a hydrogen bond, between a pair of donor−acceptor atoms present at a favorable distance.

In addition to the HB region of interest, the OD−H···OA interactions also show the angular preference elsewhere. The repulsive region in red color gradually expands in size with the decrease of the OD−H···OA angle, suggesting that OD and OA atoms are more sterically inaccessible to each other when they are located at the same side of the H atom. Accordingly, the size of the green region decreased together with the OD−H···OA angle, which corresponds to the most widespread nonspecific interactions exhibiting insignificant strength.

For comparison, Figure 4B-D shows the interactions formed between some typical hydrogen bond donors and acceptors. Generally, a similar pattern was followed by these atom pairs, in which a favorable interacting region occurs in the upper-left and a repulsive platform in the left side of each map. Except for ND-NA (Figure 3D), all other pairs show a distinct hydrogen bonding region. The most preferred interacting H···OA distances (r_0'), OD−H···OA angles (θ_0), and the corresponding potential energy (A_0) of these donor−acceptor pairs are summarized in Table 2. As one can see, the optimal angle takes 180° for all cases except ND_l-NA_p, while the optimal distance is different across various pairs. Interestingly, the related A_0 has an order of OD-OA< OD-NA≈ ND-OA< ND-NA. It means that, as far as the interactions containing the same type of donor are concerned, the most favorable potential energy in terms of acceptor types are in an order of OA < NA. The other way around is true: it is also observed on the donor side the potential order of OD < ND. These results agree well with the knowledge of that an oxygen

Table 2. Optimal Interacting Distances (r_0'), Angles (θ_0), and Corresponding Statistical Potentials (A_0) of Various Donor−Acceptor Pairs Depicted in Figures 4 and 5

atom pair types	Figure	r_0' (Å)	θ_0	A_0 (kcal/mol)
OD_p-OA_l	3A	1.8	180°	−3.00
OD_l-NA_p	3B	1.8	180°	−2.29
ND_l-OA_p	3C	2.0	180°	−2.46
ND_l-NA_p	3D	2.0	175°	−1.59
NC_p-OC_l	4A	1.8	180°	−3.32
NC_p-OA_l	4B	1.8	180°	−2.02
ND_p-OC_l	4C	2.0	180°	−2.59
NC_l-OC_p	4D	1.8	180°	−3.20
NC_p-NC_l	4E	-	-	-
NC_l-NC_p	4F	-	-	-

atom is more electronegative than a nitrogen atom, and therefore an oxygen atom as donor or acceptor leads to a stronger hydrogen bond than nitrogen. It should be noted that all the selected interaction maps shown in Figure 4 are the potentials of ligand donor and protein acceptor atom pairs, i.e. $A_{ji}(r',\theta)$, due to their higher occurrence rates in the knowledge base (Table 1). Similar trends were observed for the counterpart maps (of protein donor and ligand acceptor, $A_{ij}(r',\theta)$), though the corresponding optimum values of each map may vary.

In Figure 4, we only list a few maps of "neutral" atom pairs, i.e. atoms with a formal charge of zero. For "ionic" atom pairs, the interactions are more complicated. Take the salt bridge of NC-OC as an example: it involves not only the hydrogen bonding between the donor−acceptor pair, but also the electrostatic interaction between two charged centers. This type of interaction is also sometimes considered as a special form of particularly strong hydrogen bonds. Due to the numerous ionizible side chains of amino acids found throughout a protein, the "ionic" hydrogen bonds are widespread in protein−ligand binding interface, and many scoring functions treat the ionic and nonionic hydrogen bonds separately.[33,34] Therefore, it would be interesting to investigate the potential map of charged donor-ligand atom pairs. Figure 5 presents a few such examples: 5A and 5D show the atom pairs with opposite charges, NC-OC; 5B and 5C show the pairs comprising only one charged atom, NC-OA and ND-OC, respectively; 5E and 5F show the pairs of two positively charged atoms, NC-NC. As expected, distinct hydrogen bonding regions can be found in the first four maps but not in the last two. For 4E and 4F, since the atom of NC type is not a valid hydrogen bond acceptor (there is no lone pair in the positive charged nitrogen), no hydrogen bonds are observed in the upper left region. Instead, the hydrogen bonding region shows unfavorable interactions due to the repulsive electrostatic force between the two positively charged atoms. The most preferred geometric parameters and the corresponding potential energy of these ionic pairs are also summarized in Table 2 for comparison. Among 5A-D, two NC-OC potentials are deeper than NC-OA and ND-OC, suggesting that hydrogen bonding can be strengthened by the attractive electrostatic interactions. However, this effect only applies to the complementary ion pairs: For the atom pairs with single charged center, NC-OA and ND-OC, the most preferred potentials are −2.02 and −2.59 kcal/mol, respectively, which are at the same level of neutral atom pairs.

In addition to the optimum hydrogen bonding geometries, there are several notable properties of the potentials that can be observed from the Figure 5. For the atom pairs with opposite charges, the favorable interaction regions are not restricted to the upper left region of each map. We may observe (in 5A and 5D) that the light blue color covers large areas extending from the distance 2 Å to 6 Å. At some specific angles (θ = 20°, 50°, and 130°), the interactions are even lower than the threshold energy for hydrogen bonding. Obviously, these extensive favorable areas represent the nondirectional electrostatic interactions, instead of hydrogen bonding interactions. Another interesting finding about the composite effect is the energy barriers behind the hydrogen bonding region. For most of the donor−acceptor pairs comprising only one or no charged atoms (in either Figure 4 or 5), we see that the hydrogen bonding region is always accompanied by one or two energy barriers of 3 kcal/mol at 3−4 Å. Similar patterns were reported by Muegge and Martin in their distance-dependent PMF studies.[27] As pointed out, the apparent barriers next to the favorable interaction region reflect the preference for a hydrogen bond at a certain distance, which leaves few observed interactions at slightly longer distances. On the other hand, for the coupled ionic donor−acceptor pairs (Figure 5A and 5D), we find no such barriers in the potential, which may be offset by the favorable electrostatic attraction in the distance range of 3−4 Å. These observations support the empirical knowledge that electrostatic interactions are not as sensitive to distances or angles compared to hydrogen bonds.

4. Assessment of 2D HB Potentials in Scoring. Table 3 lists the correlation coefficients between the experimental binding affinities and the calculated energy scores of the hybrid PMF on the two test sets. It can be seen that the score obtains good correlations (R_s ranging from 0.55 to 0.61), suggesting its robustness in affinity prediction. For comparison, the results of PMF99, PMF04, GLIDE, LUDI and the 1D PMF are also presented. As described in the Methods section, Muegge's approach was followed to develop the distance-dependent potentials of this study, thus the performance difference between PMF99/04 and our 1D PMF can be mainly deemed as the dependence of statistical potentials on calibration data sets. In contrast, the difference between the hybrid and 1D PMF reflect the effects of the newly devised 2D interaction model. Clearly, the 1D PMF yields higher R_s values than PMF99 and PMF04. Since the 1D PMF scoring has been generated using 44,585 protein−ligand pairs as knowledge base, much larger than that of PMF99/04, the higher correlations of 1D PMF suggest that the increased statistical basis may achieve improved accuracy in binding affinity prediction.[27] However, the hybrid PMF does not show significant improvements over the 1D PMF: for Wang's set, the hybrid PMF gives the best correlation coefficient of 0.61, which is close to that of the 1D PMF (R = 0.60); For the core set, it shows a slightly inferior performance with a correlation coefficient of 0.55. This result is not unexpected, as the hybrid PMF is still a considerably simplified model on the basis of static crystal structures. For both hybrid and 1D PMF scores, a large part of the thermodynamic cycle represented by a binding free energy is missing, such as solvation processes and conformational changes.[2] The similar level of performances suggests that accounting for the directionality of hydrogen bond is not a determinant factor for improving the binding affinity prediction of current scoring functions. In fact, we have been aware of it in our previous study, where the superiority of GLIDE and LUDI, two scoring functions with better representation of hydrogen bond, is mainly reflected in their higher performances for binding mode identification.[23]

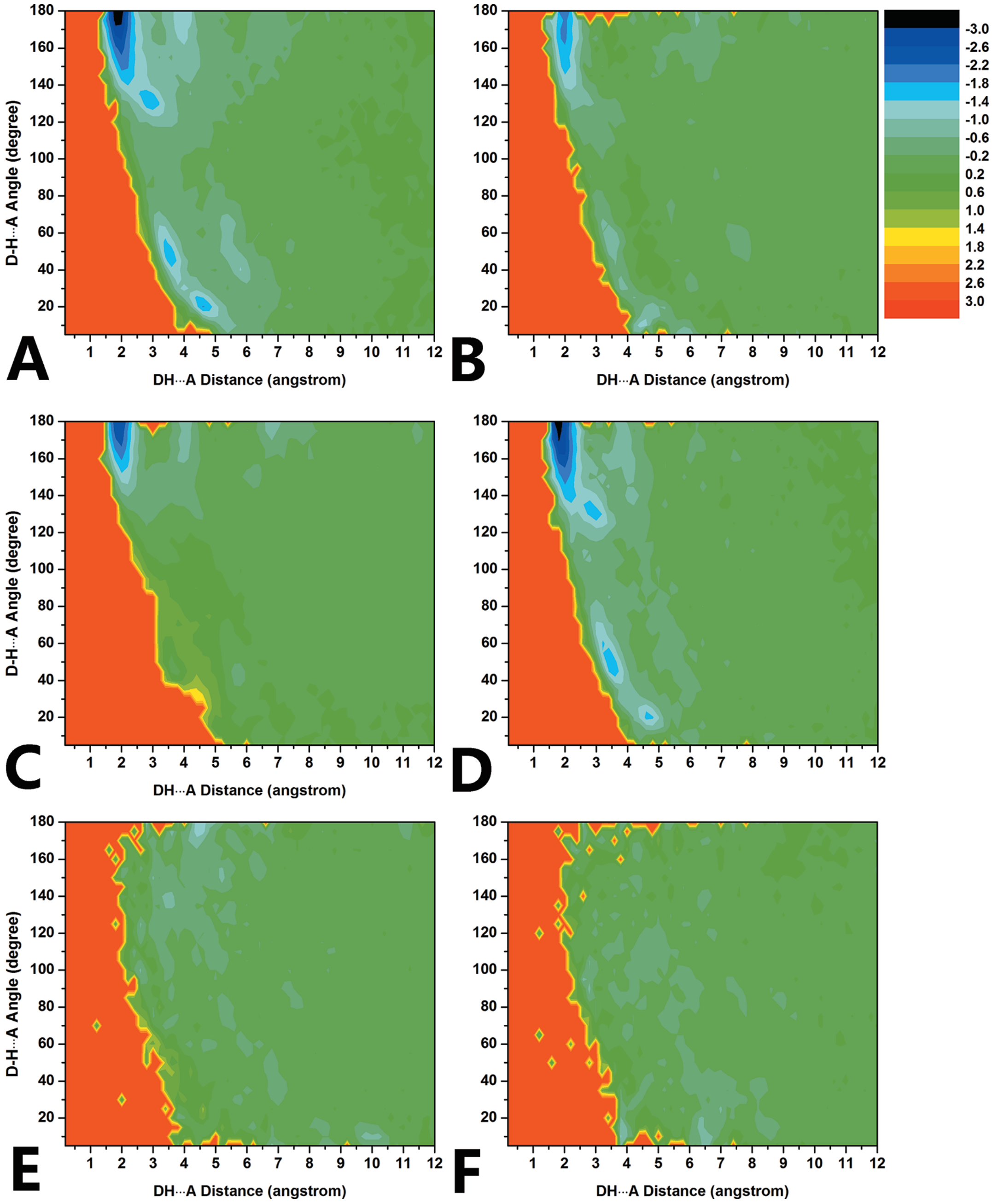


Figure 5. Heat maps of 2D HB PMF energy for a few donor−acceptor atom pairs involving ionic interactions: (**A**) NC$_p$-OC$_l$, (**B**) NC$_p$-OA$_l$, (**C**) ND$_p$-OC$_l$; (**D**) NC$_l$-OC$_p$, (**E**) NC$_p$-NC$_l$, and (**F**) NC$_l$-NC$_p$. The letter coding follows the definition given in the caption of Figure 4.

It is interesting to decompose the energy terms and examine the contributions of hydrogen bonding calculated through different interaction models. Two plots comparing the hybrid PMF score to the 1D PMF score of the hydrogen bonding interactions in the two data sets are depicted. As shown in Figure 6, significant correlations are observed for these two methods: the Pearson's correlation coefficients are 0.96 for the core set and 0.95 for Wang's set. It partly explains why the two methods show parallel results in binding affinity prediction. Intuitively, one would argue against the need for introducing an angular parameter in describing hydrogen bonds, since the obtained hydrogen bonding scores are highly correlated, and the hybrid PMF does not yield a higher predictive power of binding affinity. However, we can notice from both plots that the variance of the hybrid PMF is much larger than that of the 1D PMF: the hybrid PMF score ranges from −160 to zero kcal/mol, while the 1D PMF score is all above −60 kcal/mol. The energetic contribution of hydrogen bonding calculated from the hybrid PMF score is also much more significant: For the core set, the averaged ratio of hydrogen bonding to total interaction score is 9.1% for the 1D PMF and 18.1% for the hybrid PMF; for Wang's set, the ratios for the 1D and hybrid PMF are 11.3% and 21.8%, respectively. These results indicate that the 2D interaction model of the hybrid PMF is more sensitive to the change of hydrogen bonding, compared to the conventional 1D model. A protein−ligand system involving extensive hydrogen bonding interactions would benefit from the newly designed potentials due to the decreased score. A closer examination of Figure 6 reveals that data points are noticeably more scattered in the tight hydrogen bonding region, suggesting that the distance-dependent model is not a good proxy for the 2D model for the corresponding complexes.

The hybrid PMF based on the 2D interaction model was also tested for identifying the native or near-native binding poses. As described above, for each protein−ligand complex in test sets, twenty-one discrete binding poses including one from the crystal structure and the others generated by the program AutoDock Vina were reranked according to their hybrid PMF scores. For the core set, the experimentally determined ligand geometries of 51.6% of all cases, i.e. 113 out of 219 evaluated complexes, are ranked best out of the decoy set; for Wang's set, 62 out of 100 complexes are correctly identified. In comparison, the recognition rates of the 1D PMF for the core set and Wang's set are 34.7% and 47.0%, respectively; the rates of Vina score for these two sets are 1.8% and 4.0%. Clearly, the hybrid PMF shows better performances in the binding pose reranking test, justifying the usefulness of the 2D model in describing hydrogen bonding interactions.

Our reranking protocol also allows for the test of retrieving the near native pose out of the same pool of decoys. For the core set, setting the threshold for a valid structure to a rmsd value of 2.0 Å, we obtained that the success rates of the hybrid PMF, 1D PMF are 71.2% and 62.6%, respectively; for Wang's set, the corresponding rates are 79.0% and 73.0%. As shown in Figure 7, we can clearly observe that the success rates of the hybrid PMF are consistently higher than those of 1D PMF and Vina score at every rmsd level. This result demonstrates that application of 2D interaction model yields an essential improvement in identifying native or "well-docked" ligand poses, compared to the distance-dependent scoring functions. In our previous study, we have discussed potential reasons why current knowledge-based scoring functions show less satisfying performances in the binding-pose reranking test.[23] The current study evidently confirms our argument that consideration of directionality of atom pair

Table 3. Correlations between Experimentally Determined Binding Affinities and Calculated Binding Scores

data set	scoring function	R_s
PDBBind core set	PMF99	0.39[a]
	PMF04	0.18
	1D PMF	0.58
	GLIDE	0.44[a]
	LUDI	0.49[a]
	Hybrid PMF	0.55
Wang's set	PMF99	0.37[b]
	PMF04	0.37
	GLIDE	NA
	LUDI	0.43[b]
	1D PMF	0.60
	Hybrid PMF	0.61

[a] The results were obtained from ref 23. [b] The result was obtained from ref 32.

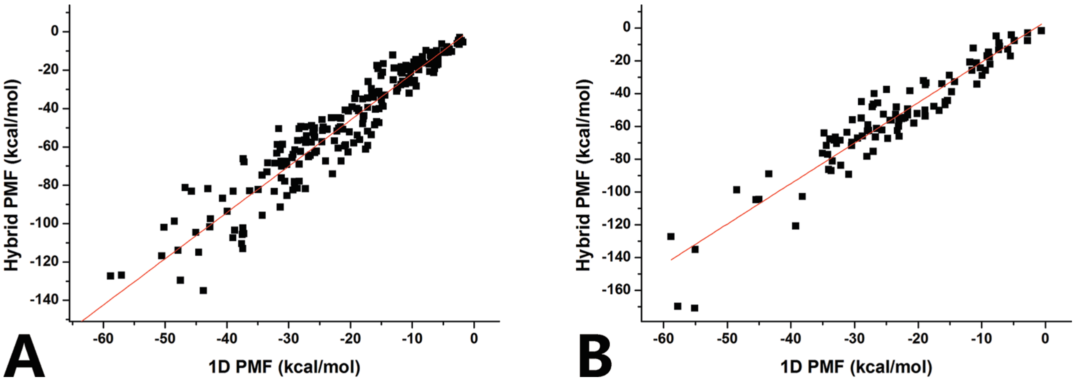


Figure 6. Scatter plots showing the HB bonding scores calculated by the 2D interaction model vs the conventional 1D interaction model for (A) the core set and (B) Wang's set.

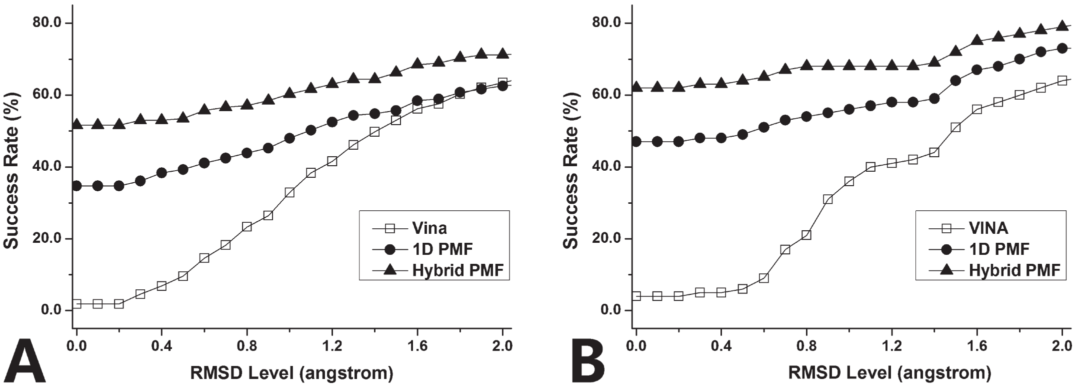


Figure 7. Cumulative plots to show the change of success rates at different rmsd levels for three types of scores for (A) the core set and (B) Wang's set.

potentials, especially the geometry of hydrogen bonds, would be crucial for scoring function improvements.

CONCLUSION

In this study, we have formulated a two-dimensional potential model describing the features of hydrogen bonding in protein—ligand complex structures and investigated its proposed applications in knowledge-based scoring function. Considering donor-ligand atoms pairs with more than 20,000 occurrences in a comprehensive knowledge base of 44,585 complexes, altogether 62 types of statistically significant 2D HB potentials have been derived. These obtained potentials clearly show that the geometry of hydrogen bond follows strict rules, either in terms of distances or angles between neighboring donor and acceptor atoms. In addition, the relative strength of some kinds of typical hydrogen bonds and their respective optimum geometrics are discussed, which agree well with empirical observations. In principle, application of the presented 2D HB potential model should benefit any knowledge-based scoring functions. However, as demonstrated by our scoring test results, it is less beneficial for the binding affinity prediction because the current distance-dependent 1D potentials are already very well at balancing many opposing contributions to binding. For the test of binding poses reranking capability, in contrast, the scoring function applying both 1D and 2D HB potentials (the hybrid PMF score) shows significant improvements over the one only applying 1D potentials. The better performance of the hybrid score confirms the advantage of the 2D model in describing hydrogen bonding interactions and also exhibits its intended usage in analyzing the binding mode of a known active ligand. The hybrid scoring approach provides a well-defined scheme to apply the 2D potentials in compatible with the use of 1D potentials, which is applicable to other knowledge-based scoring functions. The presented 2D potential model is by no means complete, whereas it may help to improve our understanding of molecular recognition in biological systems. Meanwhile, it also provides a viable means to investigate other intermolecular interactions with a dependence on orientation.

ASSOCIATED CONTENT

Ⓢ Supporting Information. A csv table listing all HB donor-ligand types and associated occurrence numbers, a text file formatted printing 62 types of 2D HB potentials, plots of a few typical 1D-potentials that may or may not form hydrogen bonding, and a Python script for assigning the protein and ligand atom types. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: 86-21-50271399. Fax: +86-21-50807088. E-mail: myzheng@mail.shcnc.ac.cn (M.Z.) or hljiang@mail.shcnc.ac.cn (H.J.).

Author Contributions

¶These authors contributed equally to this work.

ACKNOWLEDGMENT

This work is supported by the State Key Laboratory of Drug Research. The authors gratefully acknowledge financial support from the State Key Program of Basic Research of China (Grant 2009CB918502), and the National Natural Science Foundation of China (Grant 81001399 and 81072580).

REFERENCES

(1) Sarkhel, S.; Desiraju, G. R. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. *Proteins* **2004**, *54*, 247–259.

(2) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(3) Abraham, M. H.; Ibrahim, A.; Zissimos, A. M.; Zhao, Y. H.; Comer, J.; Reynolds, D. P. Application of hydrogen bonding calculations in property based drug design. *Drug Discovery Today* **2002**, *7*, 1056–1063.

(4) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777–793.

(5) Choi, H.; Kang, H.; Park, H. New angle-dependent potential energy function for backbone-backbone hydrogen bond in protein-protein interactions. *J. Comput. Chem.* **2010**, *31*, 897–903.

(6) Grzybowski, B.; Ishchenko, A.; DeWitte, R.; Whitesides, G.; Shakhnovich, E. Development of a Knowledge-Based Potential for Crystals of Small Organic Molecules: Calculation of Energy Surfaces for C=0•••H—N Hydrogen Bonds. *J. Phys. Chem. B* **2000**, *104* 7293–7298.

(7) Chen, Y.; Kortemme, T.; Robertson, T.; Baker, D.; Varani, G. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Res.* **2004**, *32*, 5147–5162.

(8) Kortemme, T.; Morozov, A. V.; Baker, D. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* **2003**, *326*, 1239–1259.

(9) Grishaev, A.; Bax, A. An empirical backbone-backbone hydrogen-bonding potential in proteins and its applications to NMR structure refinement and validation. *J. Am. Chem. Soc.* **2004**, *126*, 7281–7292.

(10) Huang, J.; Meuwly, M. Explicit Hydrogen-Bond Potentials and Their Application to NMR Scalar Couplings in Proteins. *J. Chem. Theory Comput.* **2010**, *6*, 467–476.

(11) Grabowski, S. J. What Is the Covalency of Hydrogen Bonding?. *Chem. Rev.* **2011**, *111*, 2597–2625.

(12) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 6946–6951.

(13) Liu, Z.; Wang, G.; Li, Z.; Wang, R. Geometrical Preferences of the Hydrogen Bonds on Protein—Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations. *J. Chem. Theory Comput.* **2008**, *4*, 1959–1973.

(14) Desiraju, G. R. Hydrogen bridges in crystal engineering: interactions without borders. *Acc. Chem. Res.* **2002**, *35*, 565–573.

(15) Mitchell, J. B. O.; Price, S. L. The nature of the N-H O ═ C hydrogen-bond. An intermolecular perturbation theory study of the formamide formaldehyde complex. *J. Comput. Chem.* **1990**, *11*, 1217–1233.

(16) No, K. T.; Kwon, O. Y.; Kim, S. Y.; Jhon, M. S.; Scheraga, H. A. A Simple Functional Representation of Angular-Dependent Hydrogen-Bonded Systems. 1. Amide, Carboxylic Acid, and Amide-Carboxylic Acid Pairs. *J. Phys. Chem.* **1995**, *99*, 3478–3486.

(17) Lii, J.-H.; Allinger, N. L. Directional hydrogen bonding in the MM3 force field. I. *J. Phys. Org. Chem.* **1994**, *7*, 591–609.

(18) Lii, J.-H.; Allinger, N. L. Directional hydrogen bonding in the MM3 force field: II. *J. Comput. Chem.* **1998**, *19*, 1001–1016.

(19) Buck, M.; Karplus, M. Hydrogen Bond Energetics: A Simulation and Statistical Analysis of N-Methyl Acetamide (NMA), Water, and Human Lysozyme. *J. Phys. Chem. B* **2001**, *105*, 11000–11015.

(20) 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.

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

(22) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231–235.

(23) Shen, Q.; Xiong, B.; Zheng, M.; Luo, X.; Luo, C.; Liu, X.; Du, Y.; Li, J.; Zhu, W.; Shen, J.; Jiang, H. Knowledge-based scoring functions in drug design: 2. Can the knowledge base be enriched?. *J. Chem. Inf. Model.* **2011**, *51*, 386–397.

(24) Xue, M.; Zheng, M.; Xiong, B.; Li, Y.; Jiang, H.; Shen, J. Knowledge-Based Scoring Functions in Drug Design. 1. Developing a Target-Specific Method for Kinase-Ligand Interactions. *J. Chem. Inf. Model.* **2010**, *50*, 1378–1386.

(25) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418–425.

(26) Muegge, I. PMF scoring revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(27) 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.

(28) Zhao, X. Y.; Liu, X. F.; Wang, Y. Y.; Chen, Z.; Kang, L.; Zhang, H. L.; Luo, X. M.; Zhu, W. L.; Chen, K. X.; Li, H. L.; Wang, X. C.; Jiang, H. L. An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA. *J. Chem. Inf. Model.* **2008**, *48*, 1438–1447.

(29) Rykunov, D.; Fiser, A. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinf.* **2010**, *11*, 128.

(30) Wang, R. X. A Brief Introduction to the PDBbind Database v.2009. http://www.pdbbind.org.cn (accessed Oct 1, 2010).

(31) Wang, R. X.; Fang, X. L.; Lu, Y. P.; Yang, C. Y.; Wang, S. M. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(32) 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.

(33) Bohm, 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.

(34) Bohm, 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.