JCTC Journal of Chemical Theory and Computation

Geometrical Preferences of the Hydrogen Bonds on Protein−Ligand Binding Interface Derived from Statistical Surveys and Quantum Mechanics Calculations

Zhiguo Liu, Guitao Wang, Zhanting Li, and Renxiao Wang*

State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Road, Shanghai 200032, P. R. China

Received July 8, 2008

Abstract: We have conducted potential of mean force (PMF) analyses to derive the geometrical parameters of various types of hydrogen bonds on protein−ligand binding interface. Our PMF analyses are based on a set of 4535 high-quality protein−ligand complex structures, which are compiled through a systematic mining of the entire Protein Data Bank. Hydrogen bond donor and acceptor atoms are classified into several basic types. Both distance- and angle-dependent statistical potentials are derived for each donor−acceptor pair, from which distance and angle cutoffs are obtained in an objective, unambiguous manner. These donor−acceptor pairs are also studied by quantum mechanics (QM) calculations at the MP2/6−311++G** level on model molecules. Comparison of the outcomes of PMF analyses and QM calculations suggests that QM calculation may serve as an alternative approach for characterizing hydrogen bond geometry. Both of our PMF analyses and QM calculations indicate that C−H···O hydrogen bonds are relatively weak as compared to common hydrogen bonds formed between nitrogen and oxygen atoms. A survey on the protein−ligand complex structures in our data set has revealed that C_α-H···O hydrogen bonds observed in protein−ligand binding are frequently accompanied by bifurcate N−H···O hydrogen bonds. Thus, the C_α-H···O hydrogen bonds in such cases would better be interpreted as secondary interactions.

1. Introduction

Hydrogen bonding is probably the most important factor for maintaining the molecular structures and functions of various biological as well as chemical systems.[1,2] The very basic characteristics of a hydrogen bond is the D-H···A alignment, in which the hydrogen donor (*D*) is normally a strong electronegative atom such as nitrogen or oxygen, while the hydrogen acceptor (*A*) is another electronegative atom with at least one electron lone pair. Dissociation energy of a hydrogen bond may vary from 1 kcal/mol for a weak hydrogen bond such as C−H···O to 40 kcal/mol for a strong ionic hydrogen bond such as FH···F$^-$.[3] This feature endows hydrogen bonding an essential dual role: on one hand, hydrogen bonds are relatively weak compared to covalent bonds, thus they may form and break rapidly during the process of a conformational change or molecular recognition; on the other hand, due to the considerable strength and directional nature of hydrogen bonds, a desired specificity in structure can be eventually achieved.

An in-depth understanding of protein−ligand interactions has laid the foundation of structure-based drug design techniques, such as virtual screening,[4−6] *de novo* design,[7,8] and fragment-based design.[9−11] Hydrogen bonding is an essential factor in the binding process of a ligand molecule to its target protein. Many computational studies on this subject need to detect hydrogen bonds with rule-based algorithms, which rely on interpreting the relative positions and orientations, such as the D−A distance and the D-H-A angle, of two interacting chemical groups. Such algorithms are also implemented in some empirical scoring functions, such as the ones in LUDI,[12,13] FlexX,[14] ChemScore,[15,16] GlideScore,[17] SCORE,[18] and X-Score,[19] to estimate the contribution of hydrogen bonds to protein−ligand binding affinities. Char-

* Corresponding author phone: 86-21-54925128; e-mail: wangrx@mail.sioc.ac.cn.

10.1021/ct800267x CCC: $40.75
Published on Web 10/15/2008

acterization of hydrogen bonds is also an essential factor in some other theoretical studies, such as protein folding. Therefore, deduction of the preferred geometrical parameters of various types of hydrogen bonds is a meaningful goal for all these studies.

Geometrical parameters of hydrogen bonds can be derived from a statistical survey on a large number of crystal structures. Some studies of this kind have been reported before,[20−22] which were based on either the Cambridge Structure Database[23] (CSD) or the Protein Data Bank (PDB).[24] For the purpose of characterizing the hydrogen bonds on protein−ligand binding interface, apparently the latter approach is more straightforward. Due to the rapid progress in structural biology, the total number of available three-dimensional structures of biological macromolecules is growing constantly. While this manuscript is in preparation, over 50,000 structures have already been deposited in PDB. According to our previous analyses,[25,26] up to 40% of them can be classified as valid protein−ligand complexes. High-resolution structures of these protein−ligand complexes can serve as a solid basis for conducting statistical surveys regarding the hydrogen bonds on protein−ligand binding interface.

In this study, we have applied the potential of mean force (PMF) analysis on a large number of high-quality crystal structures of protein−ligand complexes to derive the geometrical preferences of various types of hydrogen bonds. It must be mentioned that the term “potential of mean force” could be confusing since it is actually more frequently used in other areas of molecular modeling, such as the molecular dynamics simulation of liquid phases. The PMF analysis applied in our study refers to the approach proposed by Sippl et al., which was originally applied to protein folding studies.[27,28] In recent years, this approach has been extended to the evaluation of protein−ligand binding by a number of scoring functions, such as PMF-Score,[29−31] DrugScore,[32,33] BLEEP,[34,35] SMoG,[36,37] DFIRE,[38,39] and M-Score.[40] The primary aim of our study is not to develop another PMF-based scoring function. Instead, we apply this approach to the characterization of the hydrogen bonds in protein−ligand binding, which is the first of this kind to the best of our knowledge. Our study covers common hydrogen bonds formed between oxygen and nitrogen atoms as well as C−H···O hydrogen bonds. To make comparison with the outcomes of our PMF analyses, we have also employed quantum mechanics (QM) calculations on some model molecules to characterize these hydrogen bonds. The geometrical parameters of various types of hydrogen bonds deduced in our study can be readily utilized by the empirical algorithms for perceiving hydrogen bonds in protein−ligand binding or protein folding studies.

2. Computational Details

2.1. Preparation of Protein−Ligand Complex Structures. Our statistical survey is conducted on a large set of high-quality structures of protein−ligand complexes. These complexes are selected throughout the entire Protein Data Bank (PDB) through a procedure similar to the one developed by us in the compilation of the PDBbind database.[25,26] This procedure can be described briefly as following. First, the composition of protein−ligand complex is considered. PDB entries which do not contain at least one protein molecule and one valid small-molecule ligand are filtered out. Here, a valid ligand must not be a cofactor/coenzyme (such as Heme, CoA, NAD, FAD, and their derivatives) or any component of an organic solvent and buffer. It also must not contain any uncommon elements, such as Be, B, Si, and metal atoms, and its molecular weight shall not exceed 1000. Note that oligopeptides (up to 9 residues) and oligonucleotides (up to 3 residues) are considered as valid small-molecule ligands in our study. Second, the quality of protein−ligand complex structure is considered. Only the protein−ligand complex structures which are determined through crystal diffraction with an overall resolution better than or equal to 2.5 Å are accepted. Finally, each qualified complex should be formed by one protein molecule with one ligand molecule in a binary manner, i.e. there should not be multiple ligands residing in close vicinity at the same binding site. In addition, covalently bound complexes are filtered out. All of the above examinations are conducted by a set of computer programs, which make judgments based on the contents of the original structural files downloaded from PDB. Manual inspections and adjustments are also employed whenever necessary.

We have screened the entire PDB (as released in January 2006) through the above procedure, and the outcome is a list of 4535 protein−ligand complexes. The structures of all of these complexes in the PDB format are downloaded from the RCSB PDB Web site (http://www.rcsb.org/pdb/). Each complex structure is then processed into appropriate formats for the convenience of subsequent analyses. In brief, each complex is split into a ligand molecule and a complete “biological unit” of the protein molecule, and they are saved in two separated files. Other components in the original PDB file, such as water and other solvent molecules, are ignored. The protein structure is sufficiently presented by the PDB format and thus does not need any additional treatment. The ligand structure, however, needs to be interpreted properly since the atom/bond type information is largely missing in the PDB format. The I-interpret program[41] is applied here to tackle this problem. This program interprets the chemical structure of a given organic molecule with a high accuracy merely based on the identities and coordinates of its component atoms. Each processed ligand is saved in the Mol2 format and is further manually inspected in the graphical interface of the Sybyl software[42] in order to detect any remaining problems in atom/bond types.

Since the primary aim of our study is to analyze hydrogen bonds, it is necessary to specify the explicit positions of hydrogen atoms on the protein and the ligand, which are normally absent in the original PDB structural files. In our study, the “standard” protonation states under neutral pH are applied to the ligand side, i.e. carboxylic, sulfonic, and phosphoric acid groups are set in deprotonated forms, while aliphatic amine groups, guanidine, and amidine groups are set in protonated forms. Hydrogen atoms are added onto the ligand accordingly with the Sybyl software. Situations on

Table 1. Hydrogen Bond Donor and Acceptor Types Defined in Our PMF Analyses

symbol	SMARTS string	description
	Donor Types	
OD.H	[$([#8]([#1])[#6])]	*sp*^{3} oxygen atom in a hydroxyl group
ND.3	[$([#7^3][#1])]	*sp*^{3} nitrogen atom in an amine group, positively charged
ND.AM	[$([#7]([#1])[#6,#15,#16]=[#8]),$([#7]([#1])[#6]=[#16])]	nitrogen atom in an amide group
ND.PL3	[$([#7;^2;D3][#1])]	*sp*^{3} or *sp*^{2} nitrogen atom with a triangle planar geometry[a]
CD.G	[$([#6][#1])]	generic carbon atom
CD.A	[$([#6]([#7])([#6]=[#8])[#1])]	alpha-carbon on an amino acid residue[b]
	Acceptor Types	
OA.2	[$([#8]=*)]	*sp*^{2} oxygen atom
OA.H	[$([#8;D2][#1])]	*sp*^{3} oxygen atom in a hydroxyl group
OA.E	[$([#8;D2;H0])]	*sp*^{3} oxygen atom in an ester or ether group
OA.NC	[$([#8;D1]~[#6,#15,#16]~[#8;D1])]	oxygen atom in a carboxylic group, negatively charged
NA.2	[$([#7;D2])	*sp*^{2} nitrogen atom

[a] Such as the nitrogen atom in pyrrole and the one in aniline. [b] Only applicable to protein molecules.

the protein side are more complicated since the protonation status of an amino acid residue may be affected by its surrounding environment. The PROPKA algorithm[43] is employed in our study to determine the protonation status of ionizable residues under neutral pH. This algorithm is chosen since its performance was the best in a recent benchmark.[44] Hydrogen atoms are then added onto the protein structure with the AMBER program[45] according to the predictions by PROPKA.

2.2. Probing of Donor−Acceptor Pairs. An in-house C++ program, PLHB, is developed based on the open source library in OpenBabel.[46] It is used to probe the donor−acceptor pairs on the binding interface of all of the protein−ligand complexes in our data set. Donor atoms and acceptor atoms are classified into several categories according to their chemical natures (Table 1). Combination of these donor and acceptor types covers most common hydrogen bonds observed between proteins and small-molecule ligands. The SMARTS chemical language[47] and the Programmable ATom TYper (PATTY) algorithm[48] are applied to this typing scheme. With SMARTS and PATTY, flexible and efficient atom type classification can be expressed in text strings that are interpretable to chemists.

Our PLHB program also computes the desired geometrical parameters of donor−acceptor pairs, including the D−A distance (d) and the D-H-A angle (θ). Computation of the D-H-A angle needs the coordinates of hydrogen atoms, which are normally not available in the original structural files from PDB. Coordinates of the hydrogen atoms on most chemical groups can be reliably predicted with standard bond lengths, bond angles, and dihedral angles based on the hybridization state of their root atoms. An obvious exception is the hydrogen atom on a hydroxyl group (i.e., R-OH), which may have multiple possible positions around the R-O axis due to a low-energy rotation barrier. A simple searching algorithm is implemented in our PLHB program to tackle this problem: if a hydroxyl group is in close vicinity to an acceptor group on the counteracting molecule, the hydrogen atom on this hydroxyl group will be rotated around the R-O axis systematically to achieve the largest possible value of the D-H-A angle. The final coordinates of this hydrogen atom will be used in our statistical survey.

2.3. Derivation of Statistical Potentials. Pairwise potentials between donors and acceptors are derived from our data set of protein−ligand complexes through the potential of mean force (PMF) analysis. The basic idea beneath PMF analysis[27,28] is that statistically more populated configurations are energetically more favorable, and the ensemble of all accessible configurations are assumed to obey a Boltzmann distribution. In our study, the distance-dependent potential of each donor−acceptor pair is computed as

$$D_{ij}(d) = -RT\ln\left[\frac{m_0 + m_{ij}\dfrac{f_{ij}(d)}{g(d)}}{m_0 + m_{ij}}\right] \tag{1}$$

where $f_{ij}(d)$ is the relative probability of observing donor−acceptor pair i-j at distance d, and $g(d)$ is the relative probability of observing a reference state at the same distance. Since our aim is to derive the geometrical preferences of hydrogen bonds over a nonspecific reference state, a reasonable choice of the reference state is all possible atom pairs, including van der Waals pairs as well as hydrogen bond pairs.

In eq 1, $f_{ij}(d)$ is computed as

$$f_{ij}(d) = \rho_{ij}(d)/\rho_{ij}(bulk) = \left(\frac{n_{ij}(d)}{4\pi d^2\Delta d}\right)\Bigg/\left(\frac{\sum_{D_{\min}}^{D_{\max}} n_{ij}(d)}{\int_{D_{\min}}^{D_{\max}} 4\pi d^2\Delta d}\right) \tag{2}$$

And, $g(d)$ is computed as

$$g(d) = \rho_{all}(d)/\rho_{all}(bulk) = \left(\frac{n_{all}(d)}{4\pi d^2\Delta d}\right)\Bigg/\left(\frac{\sum_{D_{\min}}^{D_{\max}} n_{all}(d)}{\int_{D_{\min}}^{D_{\max}} 4\pi d^2\Delta d}\right) \tag{3}$$

Here, $\rho_{ij}(d)$ is the numerical density of donor−acceptor pair i-j observed at distance d, while $\rho_{ij}(bulk)$ is the numerical density of donor−acceptor pair i-j observed throughout the entire sampling space. $\rho_{all}(d)$ and $\rho_{all}(bulk)$ are defined similarly, which are applied to all atom pairs. In our study, the lower bound (D_{min}) and the upper bound (D_{max}) of distance cutoff are set to 2.0 Å and 8.0 Å, respectively. In order to count the occurrence (n_{ij}) of donor−acceptor pair i-j at a particular distance, the spherical sampling space centered at the donor atom is divided into multiple layers (Figure 1A). The bin width, *i.e.* Δd, is set to 0.1 Å. The

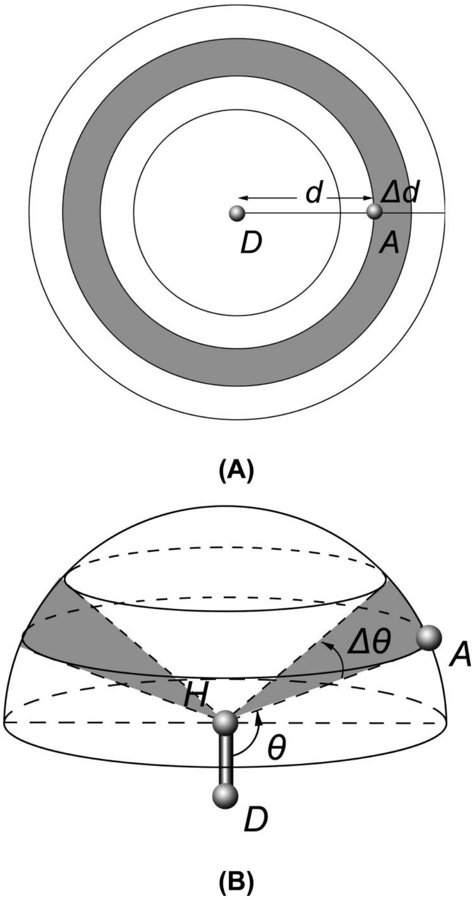

Figure 1. Geometrical parameters considered in the deduction of (*A*) distance-dependent and (*B*) angle-dependent potentials. Volume of the shaded space is computed as (*A*) $\Delta V = 4\pi d^2_{\text{D-A}}\Delta d_{\text{D-A}}$ and (*B*) $\Delta V = {}^4/_3\pi d^3_{\text{H-A}}\sin({}^{\Delta\theta}/_2)\sin(\theta+{}^{\Delta\theta}/_2)$, respectively.

equation for computing the volume of each layer is given in the legend of Figure 1.

The introduction of m_{ij} and m_0 in eq 1 is our extension to the standard algorithm for computing distance-dependent statistical potentials. m_{ij} is the total occurrence of donor−acceptor pair *i-j* within a distance cutoff of 8.0 Å. It is typical that the occurrence of donor−acceptor pair *i-j* is really low at short distance, *i.e.* $f_{ij}(d)\rightarrow 0$ and $g(d)\rightarrow 0$. A small residual m_0 is introduced so that eq 1 will produce a meaningful value close to zero in such circumstances. The value of m_0 is set to 50 in our study, which is an arbitrary choice. In fact, no noticeable difference in the outcomes of eq 1 can be observed under different values of m_0 as long as m_0 is a relatively small number (see the Supporting Information, Part 2).

In our study, angle-dependent potentials of hydrogen bonds are also derived in a similar manner. The angle-dependent potential of donor−acceptor pair *i-j* is computed as

$$A_{ij}(\theta) = -RT\ln\left[\frac{m_0 + m_{ij}\dfrac{f_{ij}(\theta)}{g_{ij}(\theta)}}{m_0 + m_{ij}}\right] \tag{4}$$

where $f_{ij}(\theta)$ is the relative probability of observing donor−acceptor pair *i-j* at a particular angle θ when a hydrogen bond between them is possible, while $g_{ij}(\theta)$ is the relative probability of observing this donor−acceptor pair at the same angle regardless if a hydrogen bond between them is possible. Since θ is only relevant to donor−acceptor pairs, the reference state in eq 4 is different from the one in eq 1.

In eq 4, $f_{ij}(\theta)$ is computed as

$$f_{ij}(\theta) = \rho_{ij}(\theta)/\rho_{ij}(bulk) = \left(\frac{n_{ij}(\theta)_{d<D_{\max}}}{\frac{4}{3}\pi d^3\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)}\right)\Bigg/\left(\frac{\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}}{\frac{2}{3}\pi d^3}\right) = \frac{n_{ij}(\theta)_{d<D_{\max}}}{2\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}} \tag{5}$$

Here $\rho_{ij}(\theta)$ is the numerical density of donor−acceptor pair *i-j* in hydrogen bonds observed at angle θ. We use a distance cutoff (D_{max}) of 3.5 Å to decide if atoms *i* and *j* are close enough to form a hydrogen bond. This cutoff is chosen since it is approximately the sum of van der Waals radii of two heavy atoms in a common N−N, N−O, or O−O hydrogen bond. $\rho_{ij}(bulk)$ is the numerical density of donor−acceptor pair *i-j* in hydrogen bonds, i.e. when $d < D_{max}$, observed throughout the entire sampling space. The lower bound (A_{min}) and the upper bound (A_{max}) of angle θ are set to 90° and 180°, respectively. In order to count the occurrence of donor−acceptor pair *i-j* at a particular angle θ, the semi-spherical sampling space centered at the hydrogen atom is divided into multiple cone-shaped sectors (Figure 1B). The bin width ($\Delta\theta$) is set to 5°. The equation for computing the volume of each sector is given in the legend of Figure 1.

In eq 4, $g_{ij}(\theta)$ is in fact computed using the same equation as $f_{ij}(\theta)$:

$$g_{ij}(\theta) = \rho_{ij}(\theta)/\rho_{ij}(bulk) = \frac{n_{ij}(\theta)_{d<D_{\max}}}{2\sin\left(\frac{\Delta\theta}{2}\right)\sin\left(\theta+\frac{\Delta\theta}{2}\right)\sum_{A_{\min}}^{A_{\max}} n_{ij}(\theta)_{d<D_{\max}}} \tag{6}$$

The only difference here is that the distance cutoff (D_{max}) is expanded to 8.0 Å. Thus, $g_{ij}(\theta)$ stands for the background probability of finding donor−acceptor pair *i-j* at angle θ regardless if they can form a valid hydrogen bond or not. m_{ij} and m_0 in eq 4 have the same meanings as in eq 1.

2.4. Quantum Mechanics Calculations. We have also applied quantum mechanics calculations on model molecules to characterize hydrogen bonds. The model molecules used in our study are shown in Figure 2, which are selected to match the donor and acceptor types considered in our PMF analyses (Table 2). These model molecules are combined to produce a total of 4 × 5 = 20 donor−acceptor complexes. An initial configuration of each donor−acceptor complex is manually constructed first, in which the D−A distance (*d*) is set to 2.5 Å, the D-H-A angle (θ) is set to 180°, and the lone pair on the acceptor atom is aligned with the A→D vector. An example is given in Figure 3A, showing how the initial configuration of the complex formed between a formylamide molecule (as the donor) and an acetone molecule (as the acceptor). The initial configuration is then subjected to structural optimization. Note that the association

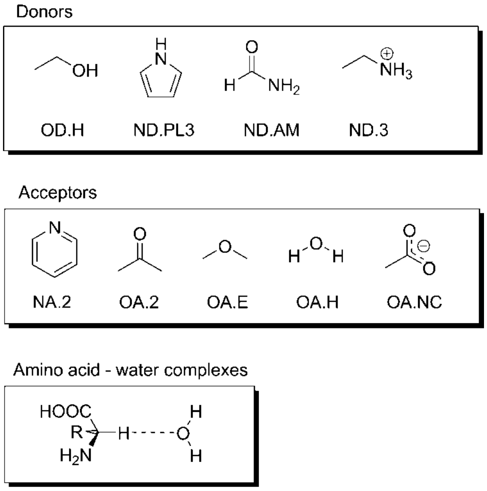

Figure 2. Model molecules used in QM calculations. The symbol below each molecule is the corresponding donor or acceptor type that it represents.

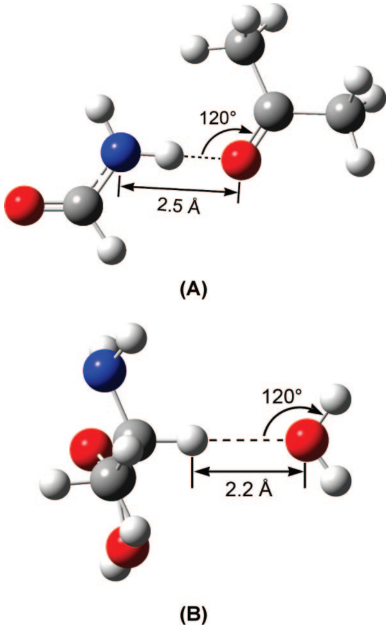

Figure 3. Illustration of the initial configuration of a donor−acceptor complex subjected to structural optimization in our QM calculation. (A) The donor is the nitrogen atom on formylamide, while the acceptor is the oxygen atom on acetone. (B) The donor is the alpha-carbon atom on alanine, while the acceptor is the oxygen atom on a water molecule.

of two model molecules involves the desired hydrogen bond as well as other secondary interactions. In order to minimize the contribution of the latter so that the overall association energy is dominated by the desired hydrogen bond, angle θ is fixed as 180° during this process. All other degrees of freedom, including d, are fully relaxed.

The optimized complex structure is then subjected to frequency analysis. The association energy of the given complex is computed as

$$\Delta E_a^{298K} = E_{D-A}^{298K} - E_D^{298K} - E_A^{298K} + E_{BSSE} \qquad (7)$$

Here, $\Delta E_{D-A}^{298\mathrm{K}}$, $\Delta E_D^{298\mathrm{K}}$, and $\Delta E_A^{298\mathrm{K}}$ are the potential energies of the complex, the donor, and the acceptor at 298 K, respectively. They all include the contributions of zero point energies and thermal energies. E_{BSSE} is the correction to the basis set superposition error (BSSE) computed with the counterpoise algorithm.[49] Based on the optimized complex structure, a potential energy scanning is also performed by varying d systematically from 2.5 to 8.0 Å at an increment of 0.1 Å. All of the rotational degrees of freedom of two molecules are fixed so that the relative orientation of two molecules does not change during this process. The association energy of any particular configuration of the given complex during potential energy scanning is also computed with eq 7. The only difference is that zero point energy and thermal energy are not computed in each case because frequency analysis on every configuration is computationally too expensive.

We have also studied the hydrogen bonds involving the alpha-carbon atoms on amino acid residues through similar QM calculations. For this purpose, model molecules of 20 natural amino acids are constructed. Each model molecule is constructed as $NH_2CHRCOOH$, in which the amino group and the carboxylic group are set in neutral forms. To simulate the protonation states of amino acid residues on protein under neutral pH condition, the side chains of Asp and Glu are set in deprotonated forms, while the side chains of Lys and Arg are set in protonated forms. A water molecule is then used as the acceptor to probe the hydrogen bonding interaction with the C_α atom on each amino acid molecule. In the initial configuration of each amino acid−water complex, the distance between the oxygen atom on the water molecule and the hydrogen atom on the C_α atom is set to 2.2 Å. The C_2 axis of the water molecule is aligned with the $C_\alpha \rightarrow H$ vector. The H-C_α-C_β-X_γ dihedral angle of each amino acid molecule is set to 180° in order to avoid steric repulsions between the side chain and the water molecule (Figure 3B). This initial configuration is the subjected to structural optimization in which the C_α-H-O angle is fixed as 180°. Computation of the association energy and the potential energy scanning for each amino acid−water complex are conducted through the same procedure described in the previous paragraph.

All calculations are performed using the GAUSSIAN 03 software[50] on an Intel Xeon 5345-based Linux cluster. Structural optimizations and single-point energy computations described above are all conducted at the MP2/6−311++G** level with frozen core approximation.

3. Results and Discussion

3.1. Geometrical Preferences of Common Hydrogen Bonds Derived from PMF Analyses. Among all of the geometrical parameters of a hydrogen bond, the D−A distance (d) and the D-H-A angle (θ) are the most widely used. The preferred values of these geometrical parameters can be derived from statistical survey on a large number of crystal structures. For example, the D−A distance associated with the highest occurrence can be considered as the optimal distance for the hydrogen bond between D and A. This is in fact the standard approach adopted by some previous studies.[20−22] Our opinion is that PMF analysis is more

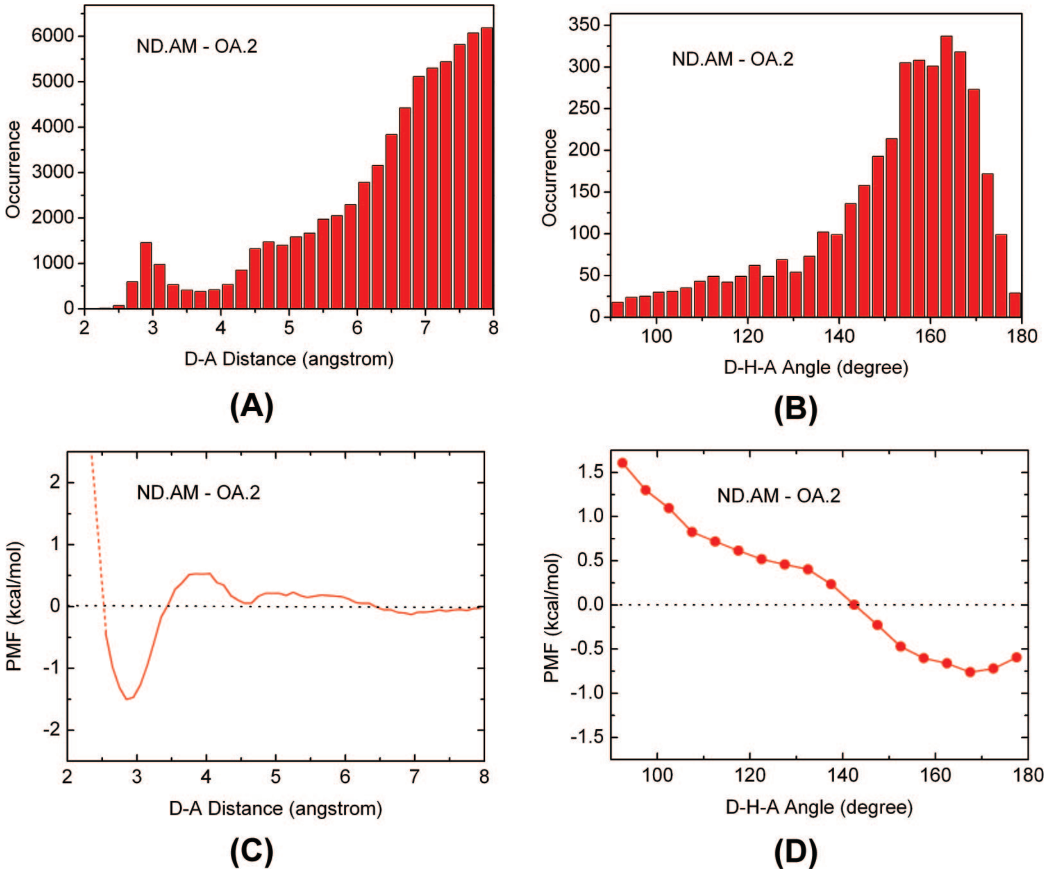

Figure 4. Distributions of (*A*) the D−A distances and (*B*) the D-H-A angles of the ND.AM-OA.2 pair observed on our data set and the corresponding (*C*) distance-dependent and (*D*) angle-dependent PMF curves of this hydrogen bonding pair.

Table 2. Optimal Interacting Distances and the Corresponding Statistical Potentials of Various Donor−Acceptor Pairs Derived from the Distance-Dependent PMF Analyses

	acceptor				
donor	NA.2	OA.2	OA.E	OA.H	OA.NC
OD.H	2.7(−1.93)[a]	2.7 (−1.44)	N/A[b]	2.8(−1.84)	2.7(−2.34)
ND.PL3	N/A	2.9 (−1.49)	3.0(−0.75)	2.9(−1.34)	2.9(−1.89)
ND.AM	3.1(−1.10)	2.9(−1.50)	3.0(−0.55)	3.0(−1.05)	2.8(−1.52)
ND.3	N/A	2.8(−1.61)	3.1(−1.70)	2.9(−2.03)	2.9(−2.33)
CD.G	N/A	3.3(−0.35)	N/A	3.3(−0.50)	3.3(−0.62)
CD.A	N/A	3.3(−0.85)	N/A	3.3(−0.24)	3.3(−0.48)

[a] The most preferred donor−acceptor interacting distance (in angstrom) of this atom pair; the number in brackets is the corresponding statistical potential (in kcal/mol) at this distance. [b] Reliable PMF curves cannot be obtained due to the low occurrence of this donor−acceptor pair in our data set.

appropriate for this purpose for two reasons. First, the optimal value of a certain geometrical parameter, e.g. the D−A distance, would be better located where the probability of finding this particular donor−acceptor pair reaches a maximum. The occurrence of this atom pair, however, does not necessarily reach its maximum at the same point. An appropriate correction is thus necessary since a larger D−A distance is associated with a larger sampling space ($\Delta V = 4\pi d^2 \Delta d$, see Figure 1A), and a larger sampling space is normally associated with higher occurrences. Second, the optimal value of a certain geometrical parameter would better be derived with consideration on its preference over an appropriate reference state. For example, when deriving the preference of the D−A distance of a hydrogen bond, it is appropriate to consider all atom pairs as the reference state (eq 1).

Here, we use a particular example, i.e. the hydrogen bonding pair ND.AM-OA.2, to further explain our approach and demonstrate its advantages. Distributions of the D−A distances and the D-H-A angles of this atom pair observed on our data set as well as the corresponding distance-dependent and angle-dependent statistical potentials derived through our approach are given in Figure 4. As one can see in Figure 4A, the occurrence of this atom pair has a local peak around 2.9 Å. This will be interpreted by a conventional counting-based approach as the optimal interacting distance (d_0) of this atom pair. Our distance-dependent PMF curve shows a sharp well at 2.9 Å, providing the same information. In addition, our PMF curve clearly shows a preferred interacting region for this atom pair, i.e. where $D_{ij}(d) < 0$ by eq 1. The upper bound of this region locates at 3.5 Å, where $D_{ij}(d) = 0$. It indicates that the hydrogen bond between

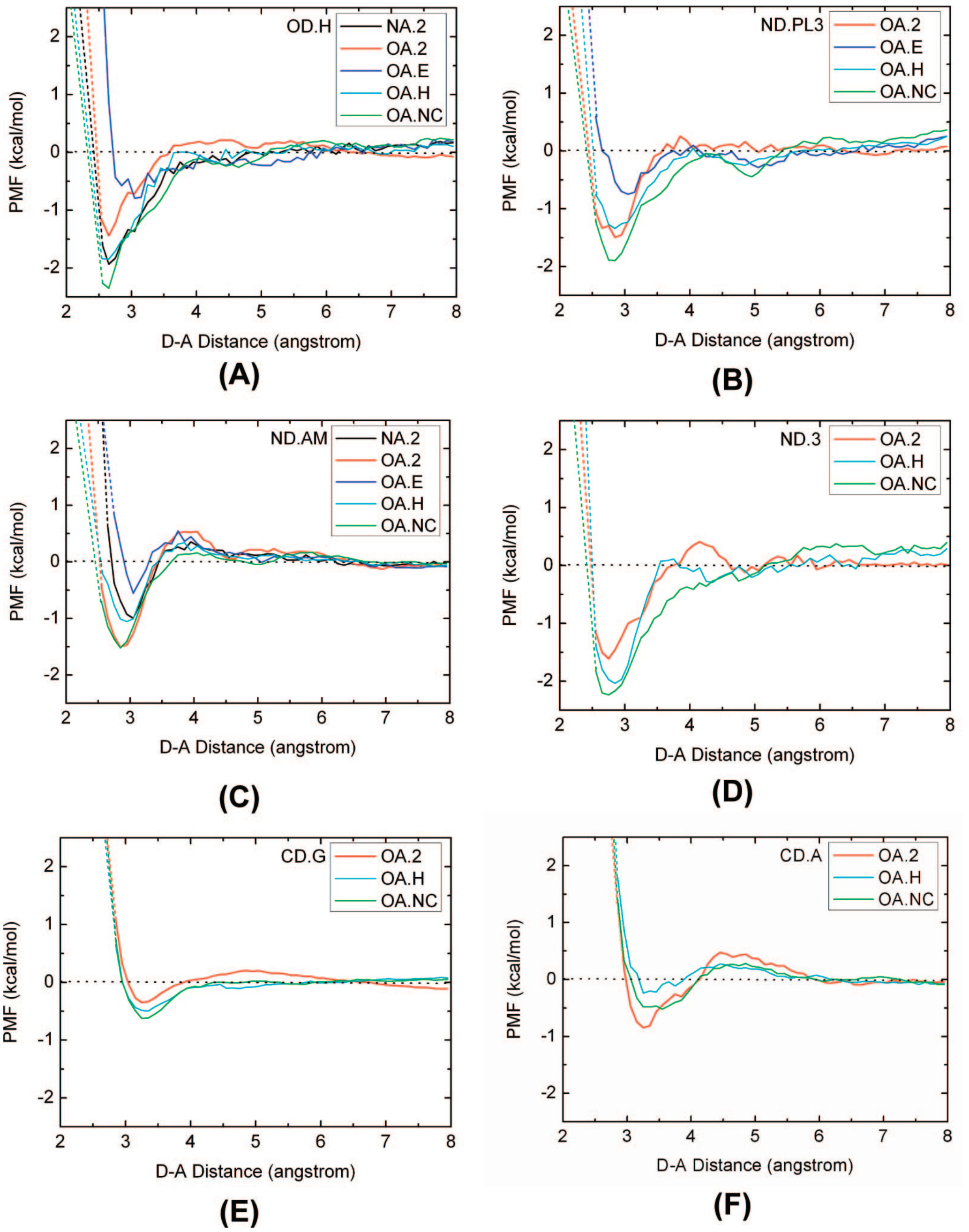

Figure 5. Distance-dependent PMF curves of donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, (*E*) CD.G, and (*F*) CD.A. These curves at short distance, e.g. < 2.5 Å, are extrapolated due to the low occurrences of donor−acceptor pairs in this range. The extrapolated segments are rendered in dashed lines.

this atom pair becomes indistinguishable from the reference state at this particular distance. This critical distance, termed as *d** in our study, can be interpreted as the distance cutoff of the given hydrogen bond. When the D−A distance is beyond this point, $D_{ij}(d)$ converges to the baseline quickly, whereas the occurrence of this atom pair keeps increasing with the D−A distance. Apparently, the conventional counting-based approach cannot deduce *d** in an unambiguous manner.

The advantage of our approach is demonstrated even more clearly in the analysis of the D-H-A angle. As one can see in Figure 4B, the occurrence of the D-H-A angle at 180° is rather low, while the highest occurrence of this angle occurs around 165°. If relying on a simple count of occurrences, one may come to the conclusion that this kind of hydrogen bond prefers a somewhat twisted geometry rather than a perfect linear alignment. This type of statement is indeed witnessed in literature from time to time. However, our study points out that the low occurrence around 180° is simply due to a much smaller sampling space at this particular angle (Figure 1B). The angle-dependent PMF curve for this atom pair actually exhibits a relatively flat bottom between 160° and 180°, still supporting the linearity assumption. The critical D-H-A angle for this donor−acceptor pair to form a valid hydrogen bond, *i.e.* θ*, can be read from this curve as 140°, where $A_{ij}(\theta) = 0$ by eq 4. At this particular angle, whether this atom pair forms a hydrogen bond or not is indistinguishable even if they are close enough to be in a

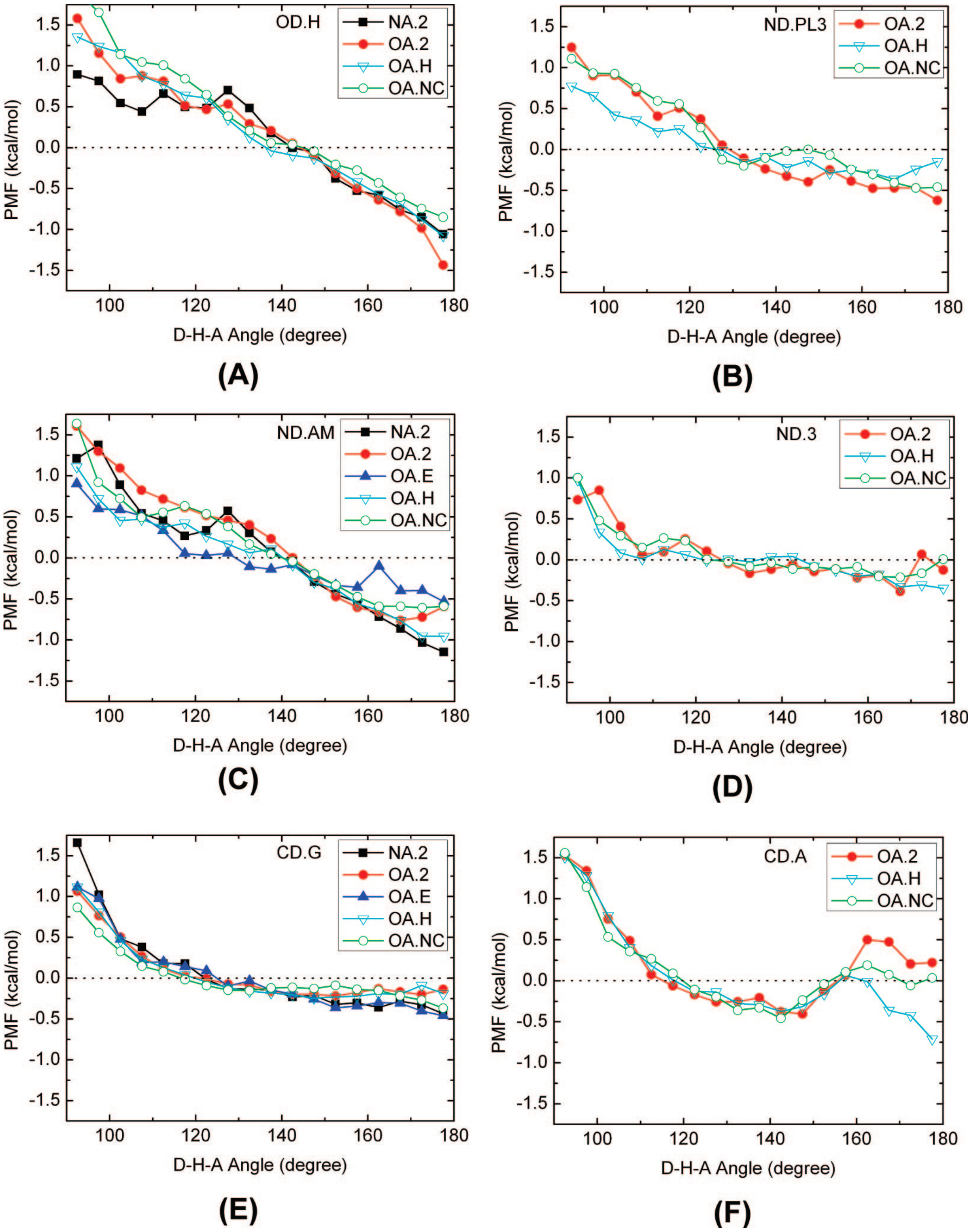

Figure 6. Angle-dependent PMF curves of donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, (*E*) CD.G, and (*F*) CD.A.

hydrogen bonding range. This angle thus can be interpreted as the angular cutoff of the given type of hydrogen bond. Again, a conventional counting-based approach cannot deduce this parameter in an unambiguous manner.

The distance-dependent PMF curves of all of the donor−acceptor pairs considered in our study are shown in Figure 5. For common hydrogen bonds formed between oxygen and nitrogen atoms, a preferred interacting region is clearly shown on each curve. The most preferred interacting distances (d_0) of all donor−acceptor pairs are summarized in Table 2. As one can see in Figure 5 and Table 2, d_0 is somewhat different across various donor−acceptor pairs. As far as the hydrogen bonds containing the same type of acceptor are concerned, d_0 values in terms of four donor types are in an order of OD.H < ND.PL3 ≈ ND.AM ≈ ND.3. This is not surprising since an oxygen atom is more electronegative than a nitrogen atom, and therefore an oxygen atom as donor leads to a shorter hydrogen bond. No obvious trend is observed on the acceptor side though. As a rule of thumb, the average d_0 values of an O−O, O−N, and N−N hydrogen bond are 2.7 Å, 2.9 Å, and 3.1 Å, respectively. Note that d_0 values of the donor−acceptor pairs containing OA.E tend to be larger by 0.1−0.2 Å than those containing other types of acceptor atoms (OA.2 or OA.H). By our definition, OA.2 or OA.H is covalently connected with only one heavy atom, while OA.E is covalently connected with two. The steric repulsion introduced by the neighboring atoms of OA.E could account for the slightly longer hydrogen bonds. As for critical distances, our results show that d^* for an O−O, O−N, and N−N hydrogen bond typically ranges from 3.5 to 4.0 Å (Figure 5). The *d** for the ND.3-OA.NC pair is significantly longer (∼5.0 Å), indicating that the interaction between these two particular atom types is electrostatic rather than a typical hydrogen bond.

Another notable feature of the distance-dependent PMF curves shown in Figure 5 is that they are not monotonous when *d* is larger than the optimal interacting distance (d_0). Instead, some fluctuations around the baseline are observed typically where $d > d^*$. Such fluctuations are the consequence of packing effects. Atoms cannot distribute freely in the three-dimensional space because they are all restricted by some chemical bonds. Moreover, each atom has a certain size so that even two unbound atoms need to be apart by a certain distance. In other words, atoms are packed in discrete layers rather than a perfect continuous manner. The same phenomenon is also seen in virtually every set of statistical potentials for protein−ligand binding derived by other researchers.[29−40]

Angular preferences are also very important for hydrogen bonds because they are directional in nature. Our definition of the angle-dependent potentials is an extension to the standard PMF approach, which is normally applied to the derivation of distance-dependent potentials. The angle-dependent PMF curves of various types of hydrogen bonds derived in our study are given in Figure 6. Our results show that all of the hydrogen bonds formed between oxygen and nitrogen atoms have a clear preference toward a linear alignment of D-H-A, *i.e.* $\theta = 180°$. Interestingly, a previous study by Desiraju et al. on a data set consisting of 28 protein−ligand complexes shows that hydrogen bonds formed between protein and ligand exhibited certain deviations from linearity.[51] In their study, a cone-correction[52] was applied to describe angular preference, an approach similar to ours. The discrepancy between our observation and theirs is probably due to the small data set employed in their study. As for the critical angle, we have observed that this parameter is basically determined by the donor type (Figure 6). As a rule of thumb, the D-H-A angular cutoffs for OD.H, ND.3, ND.PL3, and ND.AM are 140°, 125°, 125°, and 140°, respectively. These angle parameters, together with the distance parameters discussed above, can be readily utilized by empirical algorithms for perceiving hydrogen bonds.

3.2. Comparison of the Outcomes of QM Calculations and PMF Analyses. In the past two decades or so, a number of studies have employed QM calculations to characterize hydrogen bonds for various purposes.[53−59] In this study, we have followed this approach to explore the geometrical preferences of various types of hydrogen bonds. It needs to be emphasized that our QM calculations on simple model systems are independent from our PMF analyses on a large number of protein−ligand complex structures. Our purpose is to investigate if these two different approaches can achieve any consensus in terms of the geometrical and energetic properties of hydrogen bonds. One can also get a better understanding of both the strengths and shortcomings of these two approaches through this comparison.

The computed association energy of each donor−acceptor complex as a function of the D−A distance, i.e. the outcomes of potential energy scanning, is plotted in Figure 7. One can see that these distance-dependent energy curves and the corresponding distance-dependent PMF curves resemble in a qualitative manner: both types of curves exhibit a maximal interaction at a certain distance, and they converge to zero at a large distance. The optimal donor−acceptor interacting distances (d_0) of all donor−acceptor complexes are summarized in Table 3. For the neutral hydrogen bonds containing the same type of acceptors, d_0 values are in a clear order of OD.H < ND.PL3 < ND.AM, although the difference is subtle, while for the neutral hydrogen bonds containing the same type of donors, d_0 values are essentially the same across various acceptor types. As for "charged" hydrogen bonds, i.e. those containing ND.3 or OA.NC, d_0 values are generally shorter by 0.2−0.3 Å than those of neutral hydrogen bonds. Notably, the absolute values of d_0 of various hydrogen bonds given by our QM calculations agree well with those reported by Marian et al. in a recent study.[59]

Comparing the d_0 values given by PMF analyses (Table 2) with those given by QM calculations (Table 3), one can see that they match well in most cases, especially the ones associated with the same type of nitrogen donors. The discrepancy between two sets of data ranges typically between 0 and 0.3 Å (<0.1 Å for eight hydrogen bonds; 0.1−0.2 Å for four; >0.2 Å for four). Considering the limited resolution of the crystal structures in our data set, an overall agreement at this level is acceptable. Nevertheless, we have also noticed that the d_0 values of some neutral hydrogen bonds given by PMF analyses tend to be shorter by 0.1−0.3 Å than those given by QM calculations. This can be ascribed to the frequent occurrence of bifurcated hydrogen bonds in protein−ligand binding. Bifurcated hydrogen bonds are stronger intermolecular interactions and therefore are associated with shorter donor−acceptor distances as compared to single hydrogen bonds. In contrast, the model systems considered in our QM calculations allow only one hydrogen bond in each donor−acceptor complex. Thus, the results of both PMF analyses and QM calculations should be interpreted in their own contexts.

The hydrogen bonding energy is a more subtle issue. The association energies ($\Delta E^{298K}{}_{a}$) of all donor−acceptor complexes given by our QM calculations are also summarized in Table 3. The energies of the hydrogen bonds formed between neutral oxygen and nitrogen atoms range from −2.6 to −6.2 kcal/mol. The energies of the hydrogen bonds containing charged atoms (ND.3 and OA.NC) range from −15.6 to −25.3 kcal/mol, roughly 4−5 times more negative than those of neutral hydrogen bonds. This is understandable since electrostatic interactions are much more significant in the cases of charged hydrogen bonds. In contrast, the interaction potentials read from the distance-dependent PMF curves of various donor−acceptor pairs scatter in a relatively narrow range, i.e. −0.5 to −2.5 kcal/mol (Table 2). No obvious correlation can be found between the energy data in Tables 2 and 3. Particularly, the interaction potentials of charged hydrogen bonds are not significantly more negative than those of neutral hydrogen bonds.

One may argue that unlike QM computed energies, statistical potentials produced by an equation like eq 1 include solvation effects implicitly since they are derived

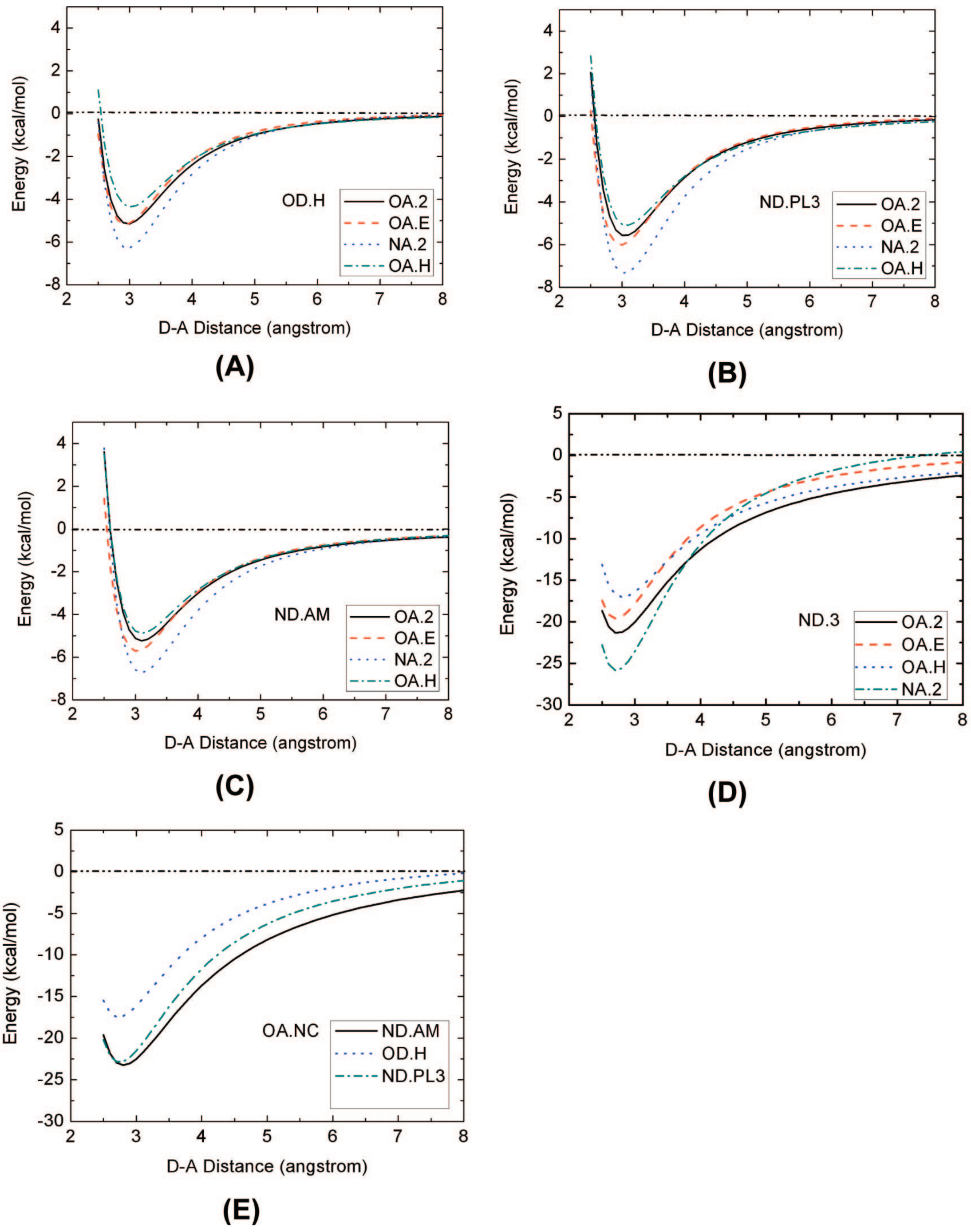

Figure 7. Association energies of the donor−acceptor complexes containing donor types (*A*) OD.H, (*B*) ND.PL3, (*C*) ND.AM, (*D*) ND.3, and acceptor type (*E*) OA.NC calculated at the MP2/6−311++G** level.

Table 3. Optimal Interacting Distances and the Corresponding Association Energies of Various Donor−Acceptor Complexes Calculated at the MP2/6−311++G** Level

	acceptor[a]				
donor[a]	NA.2	OA.2	OA.E	OA.H	OA.NC
OD.H	2.97(−4.62)[b]	2.97(−3.69)	2.93(−2.93)	3.02(−2.69)	2.73(−16.02)
ND.PL3	3.04(−6.17)	3.04(−3.40)	2.97(−4.97)	3.07(−4.13)	2.74(−22.58)
ND.AM	3.10(−4.45)	3.10(−3.02)	3.02(−4.07)	3.10(−3.99)	2.80(−21.72)
ND.3	2.72(−25.32)	2.74(−20.64)	2.71(−18.95)	2.81(−15.61)	N/A[c]

[a] The corresponding model molecules are shown in Figure 2. [b] The optimal interacting distance (in angstrom) between the donor atom and the acceptor atom; the number in brackets is the association energy of this complex at 298 K (in kcal/mol), including zero point energy and thermal energy corrections. [c] In this case, one hydrogen atom translocates from the positively charged nitrogen atom to the negatively charged oxygen atom after energy minimization of the initial configuration. The computed energy does not reflect the formation of a hydrogen bond and thus is not reported here.

from protein−ligand complex structures that are fully solvated. Formation of a charged hydrogen bond is accompanied with larger desolvation penalties, and thus the net gain is not more significant than the one of a neutral hydrogen bond. Our opinion is that this statement may not be true. Technically, eq 1 only gives the relative probability of finding a given atom pair at any particular distance. The physical basis of using the logarithm of such probability as interaction potential is actually vague. For example, one can see in Table 2 that the statistical potentials of the hydrogen bonds containing OA.E (*sp^3* oxygen atom in an ether or ester group) as acceptor are consistently less negative than those containing OA.H (*sp^3* oxygen atom in a hydroxyl group) as acceptor (−0.75 vs −1.34, −0.55 vs −1.05, and −1.70 vs −2.03). However, our QM calculations (Table 3) indicate that interaction energies of the hydrogen bonds of the former type are comparable or even more negative than those of the latter type in vacuum (−4.97 vs −4.13, −4.07 vs −3.99, and −18.95 vs −15.61). Considering that the desolvation energies of dimethyl ether (CH_3OCH_3) and methanol (CH_3OH) are 1.92 and 5.11 kcal/mol, respectively, it is not reasonable to expect that the net energy of a hydrogen bond of the former type is less negative than the counterpart of the latter type. The hydrogen bonds of the former type are associated with less negative statistical potentials simply because their occurrence is considerably lower than the one of the latter type in our data set (see the Supporting Information).

In our study, we have investigated hydrogen bonds through two different approaches. In contrast to QM calculation on simple model molecules, PMF analysis is capable of characterizing the hydrogen bonds formed during protein−ligand binding *in situ*. This approach can be used to deduce some geometrical parameters that are useful for scoring function or force field development, such as the distance and angle cutoffs of various hydrogen bonds. Such parameters are difficult to obtain through QM calculations. Nevertheless, the PMF analysis approach also has its shortcomings. One major problem is that reliable PMF potentials cannot be obtained for the atom pairs with low occurrences, e.g. those labeled as “N/A” in Table 2. Another problem is that the outcomes of PMF analysis are statistical averages, which can be ambiguous sometimes. In contrast, the outcomes of QM calculation are usually straightforward to interpret since they are obtained on idealized model systems. Also, QM calculation is technically applicable to any appropriate model systems. However, both the geometries and energies obtained through QM calculation on model molecules in vacuum need to be validated with caution in the context of protein−ligand binding. Therefore, our opinion is that PMF analysis and QM calculation are two complementary approaches to characterize the geometries of the hydrogen bonds formed in protein−ligand binding. It is however not appropriate to compare statistical potentials with QM-calculated energies since they are derived from different contexts.

3.3. On C−H···O Hydrogen Bonds. Some previous studies have reported that uncommon hydrogen bonds, especially C−H···O, are observed in the crystal structures of small organic molecules[60,61] and proteins[62−64] as well as protein−ligand and protein−protein complexes.[65−67] Some researchers state that the frequent occurrence of C−H···O hydrogen bonds indicates their essential role in the stability of protein structures.[68,69] A number of in-depth QM studies on C−H···O═C interactions have been done by Dixon et al.,[70−73] who demonstrate that such interactions can be fairly strong. However, some other studies provide conflicting conclusions that C−H···O hydrogen bonds may not make significant contributions to the stability of protein structures.[74,75] Conflicting viewpoints on the role of C−H···O hydrogen bonds in protein−ligand binding can also be found in the literature.[65,76,77] In this study, we have performed both PMF analyses and QM calculations on the C−H···O hydrogen bonds in order to investigate their elusive role.

As one can see from Figure 5E,F, the distance-dependent PMF curves of carbon donors, i.e. CD.G and CD.A, also exhibit a preferred interacting region, a feature similar to those of nitrogen and oxygen donors. However, the potential wells observed on these PMF curves are generally shallower than those of common hydrogen bonds. The most preferred interacting distance (d_0) involving carbon donors is considerably larger (∼3.3 Å) (Table 2), which agrees well with a previous survey of hydrogen bonds in protein−protein interaction.[67] Note that this distance is actually close to the sum of van der Waals radii of carbon and oxygen. The angle preferences of C−H···O hydrogen bonds are also different from those of common hydrogen bonds. The angle-dependent PMF curve of donor type CD.G is somewhat flat between 140° and 180° (Figure 6E). No strong preference toward a linear alignment of C−H···O is observed. All of these observations suggest that the C−H···O bonds on protein−ligand binding interface are quite different from common hydrogen bonds. They resemble nonspecific van der Waals contacts more closely from a statistical point of view.

We have paid special attention to the alpha-carbons on amino acid residues since they are more acidic than general carbon atoms and thus are more likely to form genuine C−H···O hydrogen bonds. Unexpectedly, the C_α-H-O angle shows a clear preference to 140° as indicated by the angle-dependent PMF curve of donor type CD.A (Figure 6F). We thus suspect that in such a case, the acceptor atom on the ligand side may form a bifurcated hydrogen bond with a nearby amide group on the protein backbone, which consequently forces the C_α-H-O angle to deviate from linearity. In order to prove this, we have re-examined the entire data set for all of the amino acid residues involved in C_α-H···O hydrogen bonds with the ligand side. Distribution of the H-$C_{i,\alpha}$-C_i-N_{i+1} dihedral angles of these residues reveals that the most populated value of this angle is around 30° (Figure 8A), an angle quite suitable for the proximity of the hydrogen atoms on $C_{i,\alpha}$ and N_{i+1}. Another piece of supportive evidence comes from the Ramachandran plot of these residues (Figure 8B). One can see that most of these residues reside in *β*-strands, which have appropriate Ψ dihedral angles facilitating the access to the hydrogen atoms on $C_{i,\alpha}$ and N_{i+1} by the same acceptor atom on the ligand side. As a matter of fact, among all the $C_{i,\alpha}$-H···O bonds observed in our data set, over 73% of them (989 in 1351) are found to be accompanied with a bifurcated N_{i+1}-H···O hydrogen bond. Considering that C−H···O hydrogen bonds are generally

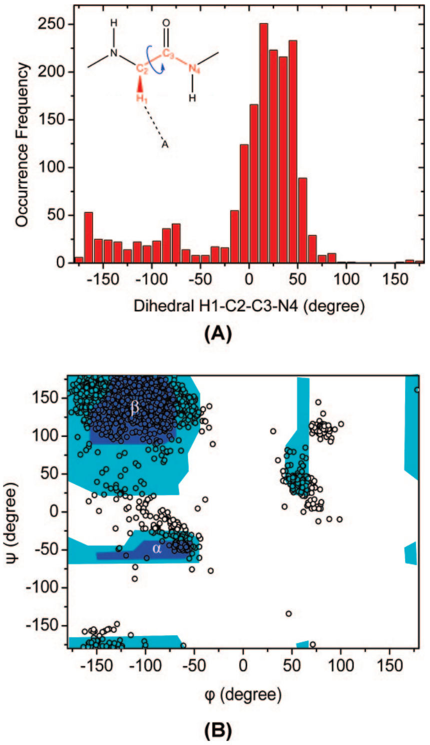

Figure 8. (*A*) Distribution of the H_1-C_2-C_3-N_4 dihedral angles of all amino acid residues which are involved in C_α-H···O hydrogen bonds. (*B*) Ramachandran plot of these residues. In both cases, a C_α-H···O hydrogen bond will be counted if the C_α-O distance is shorter than 3.6 Å and the C_α-H-O angle is larger than 90°. Glycines are excluded in this survey since they have no side chains.

much weaker than regular hydrogen bonds, it is appropriate to interpret the C_α-H···O hydrogen bond as a resultant phenomenon in such a case. It should be emphasized that our analysis does not necessarily rule out the contribution of all C−H···O bonds. For example, when a binding pocket is predominantly hydrophobic, even one C−H···O hydrogen bond can be critical for achieving the specific binding of a ligand.[78] Our analysis prompts that one needs to be extremely careful when interpreting the role of a C−H···O bond. For example, if ignoring the frequent co-occurrence of $C_{i,\alpha}$-H···O and N_{i+1}-H···O hydrogen bonds, one may come down to the wrong conclusion that a C_α-H···O hydrogen bond is an independent factor common in protein−ligand binding.

In our QM study of C_α-H···O hydrogen bonds, a water molecule is used as a probe to interact with the C_α atoms on all 20 types of amino acids. The optimal interacting distance and the corresponding association energy of each amino acid−water complex are summarized in Table 4. The optimal C_α···O interacting distances typically range from 3.5 to 3.6 Å for most amino acids. Notably, these values are again greater by ~0.2 Å than the counterparts derived from our PMF analyses. This discrepancy is understandable since there is no bifurcate N_{i+1}-H···O hydrogen bond to bring the water molecule closer to the C_α atom in the model systems considered in our QM calculations, whereas it is a frequently occurring event on real protein−ligand complexes.

Table 4. Properties of the C_α-H···O Hydrogen Bonds between 20 Amino Acids and Water Molecules Calculated at the MP2/6−311++G** Level

model	$d_{O\cdots H}$ (Å)[a]	$d_{O\cdots C}$ (Å)[b]	$\triangle d_{C-H}$ (Å)[c]	$\triangle E_a^{298K}$ (kcal/mol)[d]	$\triangle E_a^{298K}$ (kcal/mol)[e]
GLY	2.50	3.60	−0.0023	−1.70	−1.00
ALA	2.45	3.55	−0.0028	−1.53	−1.41
VAL	2.49	3.58	−0.0033	−1.81	−1.13
LEU	2.49	3.59	−0.0029	−1.65	−1.55
ILE	2.44	3.53	−0.0016	−1.86	−1.79
PHE[f]	2.43	3.52	−0.0014	−1.78	N/A
TYR[f]	2.43	3.52	−0.0014	−1.75	N/A
TRP[f]	2.44	3.54	−0.0013	−1.66	N/A
CYS	2.45	3.54	−0.0030	−2.16	−2.03
MET	2.46	3.56	−0.0032	−1.91	−1.81
ASN	2.43	3.53	−0.0025	−2.41	−1.68
GLN	2.47	3.57	−0.0016	−1.53	−0.77
SER	2.46	3.55	−0.0010	−1.82	−0.52
THR	2.46	3.56	−0.0017	−2.19	−1.45
HIS	2.37	3.46	−0.0021	−2.64	−1.91
ARG[f]	2.36	3.46	−0.0024	−4.27	N/A
LYS	2.37	3.46	−0.0028	−4.14	−2.79
ASP	2.94	4.05	0.0002	−3.39	−3.37
GLU	3.04	4.14	0.0004	−2.74	−2.74

[a] Distance between the oxygen atom on water and the hydrogen atom on the C_α atom. [b] Distance between the oxygen atom on water and the C_α atom. [c] Change in the C_α-H bond length upon the formation of the C_α-H···O hydrogen bond. [d] Association energy of the amino acid−water complex at 0 K. [e] Association energy of the amino acid−water complex at 298 K, including zero point energy and thermal energy corrections. [f] Complete convergence is not achieved in structural optimization due to the complexity of the model system. Consequently, zero point energy and thermal energy corrections are not computed since the frequency analysis is not feasible in this case.

The association energies of amino acid−water complexes produced by our QM calculations vary significantly among different types of amino acids (Table 4). As for the complexes involving neutral amino acids, the $\Delta E^{298K}{}_a$ values range between −0.5 and −2.0 kcal/mol. In this regard, these C_α-H···O hydrogen bonds are generally weaker as compared to the common hydrogen bonds formed by nitrogen and oxygen atoms (Table 3). In contrast, the complexes involving charged amino acids, including Arg, Lys, Asp, and Glu, have considerably more negative $\Delta E^{298K}{}_a$ values between −2.7 and −3.4 kcal/mol, close to the level of common hydrogen bonds. These enhanced association energies should be attributed to the charge-dipole interactions between these amino acids and water molecules. The association energy is plotted in Figure 9 as a function of C_α···O distance for three selected amino acids, i.e. Ala, Thr, and Lys. One can see that as the C_α···O distance increases, the association energy of the Lys-water complex converges to the baseline much slower than that of the Ala-water or Thr-water complex. This is a typical characteristic of long-range electrostatic interactions. Moreover, in the cases of negatively charged amino acids, i.e. Asp and Glu, the water molecule actually turns over after structural optimization, pointing its electron lone pairs rather than hydrogen atoms to the amino acid. The hydrogen atoms represent the positive end of the water dipole, and thus turning over of the water molecule will facilitate its charge-dipole interaction with the amino acid. Turning over of the water molecule, of course, eliminates the possibility of forming the C_α-H···O bond. Based on these

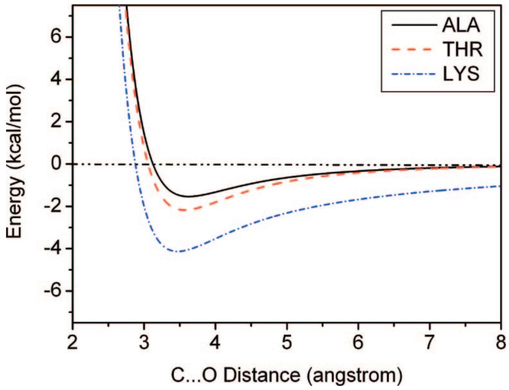

Figure 9. Association energies of three selected amino acid−water complexes calculated at the MP2/6−311++G** level.

observations, we conclude that the interactions between charged amino acids and water molecules are dominated by electrostatic interactions rather than the C_α-H···O hydrogen bond. The contribution of a C−H···O hydrogen bond in these cases may not be as significant as what some previous studies have suggested.[79]

4. Conclusions

We have analyzed the geometrical preferences of various types of hydrogen bonds found on protein−ligand binding interface. Our PMF analyses are based on a large set of high-quality protein−ligand complex structures, which is compiled through a systematic mining of the entire PDB. We have demonstrated that one can obtain both distance- and angle-dependent statistical potentials for a given type of hydrogen bond, from which distance and angle cutoffs can be obtained in an objective, unambiguous manner. Such geometrical parameters can be readily utilized by empirical algorithms for perceiving hydrogen bonds. The results given by our PMF analyses are also compared with those given by QM calculations on model molecules. The optimal interacting distances given by the two approaches are basically in accordance with each other except for a few cases. This suggests that QM calculation may serve as an alternative approach for characterizing hydrogen bond geometry especially when PMF analysis is not applicable. Nevertheless, no obvious correlation has been observed between the statistical potentials given by PMF analyses and the association energies given by QM calculations. It is not appropriate to validate QM energies with statistical potentials and *vice versa*. Both of our PMF analyses and QM calculations indicate that C−H···O hydrogen bonds are relatively weak as compared to common hydrogen bonds formed between nitrogen and oxygen atoms. In particular, our survey on protein−ligand complex structures reveals that the relatively frequent occurrence of C_α-H···O hydrogen bonds in protein−ligand binding is largely due to the coexistence of bifurcate N−H···O hydrogen bonds. Thus, the C_α-H···O hydrogen bonds in such cases would be better interpreted as secondary interactions.

Acknowledgment. The authors are grateful for the financial support from the Chinese National Natural Science Foundation (Grants 20502031 & 20772149), the Chinese Ministry of Science and Technology (the 863 high-tech project, Grant 2006AA02Z337), and the Science and Technology Commission of Shanghai Municipality (Grants 06PJ14115 & 074319113).

Supporting Information Available: Some raw data of the PMF analyses and QM calculations conducted in this study as well as more descriptions and discussion. This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Wormer, P. E.; van Der Avoird, A. *Chem. Rev.* **2000**, *100*, 4109–4144.

(2) Hobza, P.; Havlas, Z. *Chem. Rev.* **2000**, *100*, 4253–4264.

(3) Steiner, T. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 49–76.

(4) Ghosh, S.; Nie, A.; An, J.; Huang, Z. *Curr. Opin. Chem. Biol.* **2006**, *10*, 194–202.

(5) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(6) Jain, A. N. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 396–403.

(7) Schneider, G.; Fechner, U. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.

(8) Dean, P. M.; Lloyd, D. G.; Todorov, N. P. *Curr. Opin. Drug Discovery. Dev.* **2004**, *7*, 347–353.

(9) Hajduk, P. J.; Greer, J. *Nat. Rev. Drug Discovery* **2007**, *6*, 211–219.

(10) Ciulli, A.; Abell, C. *Curr. Opin. Biotechnol.* **2007**, *18*, 489–496.

(11) Verdonk, M. L.; Hartshorn, M. J. *Curr. Opin. Drug Discovery Dev.* **2004**, *7*, 404–410.

(12) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243–256.

(13) Bohm, H. J. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309–323.

(14) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J. Mol. Biol.* **1996**, *261*, 470–489.

(15) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(16) Murray, C. W.; Auton, T. R.; Eldridge, M. D. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503–519.

(17) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(18) Pei, J.; Wang, Q.; Zhou, J.; Lai, L. *Proteins* **2004**, *57*, 651–664.

(19) Wang, R.; Lai, L.; Wang, S. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 11–26.

(20) Taylor, R.; Kennard, O. *Acc. Chem. Res.* **1984**, *17*, 320–326.

(21) Mills, J. E. J.; Dean, P. M. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 607–622.

(22) Bruno, I. J.; Cole, J. C.; Lommerse, J. P. M.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525–537.

(23) Allen, F. H. *Acta. Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 380–388.

(24) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(25) Wang, R.; Fang, X.; Lu, Y.; Wang, S. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(26) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(27) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.

(28) Sippl, M. J. *Curr. Opin. Struct. Biol.* **1995**, *5*, 229–235.

(29) Muegge, I.; Martin, Y. C. *J. Med. Chem.* **1999**, *42*, 791–804.

(30) Muegge, I. *J. Comput. Chem.* **2001**, *22*, 418–425.

(31) Muegge, I. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(32) Gohlke, H.; Hendlich, M.; Klebe, G. *J. Mol. Biol.* **2000**, *295*, 337–356.

(33) Velec, H. F.; Gohlke, H.; Klebe, G. *J. Med. Chem.* **2005**, *48*, 6296–6303.

(34) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. *J. Comput. Chem.* **1999**, *20*, 1165–1176.

(35) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. *J. Comput. Chem.* **1999**, *20*, 1177–1185.

(36) Ishchenko, A. V.; Shakhnovich, E. I. *J. Med. Chem.* **2002**, *45*, 2770–2780.

(37) Dominy, B. N.; Shakhnovich, E. I. *J. Med. Chem.* **2004**, *47*, 4538–4558.

(38) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. *J. Med. Chem.* **2005**, *48*, 2325–2335.

(39) Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y. *Proteins* **2004**, *56*, 93–101.

(40) Yang, C. Y.; Wang, R.; Wang, S. *J. Med. Chem.* **2006**, *49*, 5903–5911.

(41) Zhao, Y.; Cheng, T.; Wang, R. *J. Chem. Inf. Model* **2007**, *47*, 1379–1385.

(42) *Sybyl, version 7.3*; Tripos, Inc.: St. Louis, MO 63144, U.S.A., 2006.

(43) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(44) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. *BMC Biochem*. [Online] **2006**, *7*, Article 18. http://www.biomedcentral.com/1471−2091/7/18/ (accessed Nov 27, 2007).

(45) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.;Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.;Kollman, P. A. *AMBER, version 9.0*; University of California, San Francisco: San Francisco, CA 94158−2517, U.S.A., 2004.

(46) *OpenBabel, version 2.1.0*; The Open Source Chemistry Toolbox; SourceForge, 2007.

(47) *SMARTS*; A Language for Describing Molecular Patterns; Daylight, Inc.: Aliso Viejo, CA 92656, U.S.A., 1987.

(48) Bush, B. L.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.

(49) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(50) Frisch, M.J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G. D. S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision D.01*; Gaussian, Inc.: Wallingford, CT 06492, U.S.A., 2004.

(51) Sarkhel, S.; Desiraju, G. R. *Proteins* **2004**, *54*, 247–259.

(52) Kroon, J.; Kanters, J. A. *Nature* **1974**, *248*, 667–669.

(53) Scheiner, S. *Reviews in Computational Chemistry, Calculating the Properties of Hydrogen Bonds by Ab Initio Methods*; Wiley: Hoboken, NJ 07030, U.S.A., 1991; Vol. 2.

(54) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. *J. Phys. Chem. A* **1998**, *102*, 3782–3797.

(55) Lukin, O.; Leszczynski, J. *J. Phys. Chem. A* **2002**, *106*, 6775–6782.

(56) Ireta, J.; Neugebauer, J.; Scheffler, M. *J. Phys. Chem. A* **2004**, *108*, 5692–5698.

(57) Kone, M.; Illien, B.; Graton, J.; Laurence, C. *J. Phys. Chem. A* **2005**, *109*, 11907–11913.

(58) Hao, M. H. *J. Chem. Theory Comput.* **2006**, *2*, 863–872.

(59) Raub, S.; Marian, C. M. *J. Comput. Chem.* **2007**, *28*, 1503–1515.

(60) Desiraju, G. R. *Acc. Chem. Res.* **1991**, *24*, 290–296.

(61) Taylor, R.; Kennard, O. *J. Am. Chem. Soc.* **1982**, *104*, 5063–5070.

(62) Derewenda, Z. S.; Lee, L.; Derewenda, U. *J. Mol. Biol.* **1995**, *252*, 248–262.

(63) Klaholz, B. P.; Moras, D. *Structure* **2002**, *10*, 1197–1204.

(64) Chakrabarti, P.; Chakrabarti, S. *J. Mol. Biol.* **1998**, *284*, 867–873.

(65) Pierce, A. C.; Sandretto, K. L.; Bemis, G. W. *Proteins* **2002**, *49*, 567–576.

(66) Panigrahi, S. K.; Desiraju, G. R. *Proteins* **2007**, *67*, 128–141.

(67) Jiang, L.; Lai, L. H. *J. Biol. Chem.* **2002**, *277*, 37732–37740.

(68) Madan Babu, M.; Kumar Singh, S.; Balaram, P. *J. Mol. Biol.* **2002**, *322*, 871–880.

(69) Baures, P. W.; Wiznycia, A.; Beatty, A. M. *Bioorg. Med. Chem.* **2000**, *8*, 1599–1605.

(70) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Am. Chem. Soc.* **2000**, *122*, 4750–4755.

(71) Vargas, R.; Garza, J.; Dixon, D. A.; Hay, B. P. *J. Phys. Chem. A* **2000**, *104*, 5115–5121.

(72) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 4963–4968.

(73) Vargas, R.; Garza, J.; Friesner, R. A.; Stern, H.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2005**, *109*, 6991–6992.

(74) Novoa, J. J.; Lafuente, P.; Mota, F. *Chem. Phys. Lett.* **1998**, *290*, 519–525.

(75) Yohannan, S.; Faham, S.; Yang, D.; Grosfeld, D.; Chamberlain, A. K.; Bowie, J. U. *J. Am. Chem. Soc.* **2004**, *126*, 2284–2285.

(76) Musah, R. A.; Jensen, G. M.; Rosenfeld, R. J.; McRee, D. E.; Goodin, D. B.; Bunte, S. W. *J. Am. Chem. Soc.* **1997**, *119*, 9083–9084.

(77) Toth, G.; Bowers, S. G.; Truong, A. P.; Probst, G. *Curr. Pharm. Des.* **2007**, *13*, 3476–3493.

(78) Sola, J.; Riera, A.; Verdaguer, X.; Maestro, M. A. *J. Am. Chem. Soc.* **2005**, *127*, 13629–13633.

(79) Scheiner, S.; Kar, T.; Gu, Y. *J. Biol. Chem.* **2001**, *276*, 9832–9837.

CT800267X