JCTC Journal of Chemical Theory and Computation

PACE Force Field for Protein Simulations. 1. Full Parameterization of Version 1 and Verification

Wei Han,[†] Cheuk-Kin Wan,[†] Fan Jiang,[†,‡] and Yun-Dong Wu*[,†,‡,§]

Department of Chemistry, The Hong Kong University of Science & Technology, Clear Water Bay, Kowloon, Hong Kong, China, School of Chemical Biology and Biotechnology, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen, China, and College of Chemistry, Peking University, Beijing, China

Received June 8, 2010

Abstract: A further parametrization of a united-atom protein model coupled with coarse-grained water has been carried out to cover all amino acids (AAs). The local conformational features of each AA have been fitted on the basis of restricted coil-library statistics of high-resolution X-ray crystal structures of proteins. Potential functions were developed on the basis of combined backbone and side chain rotamer conformational preferences, or rotamer Ramachandran plots (ϕ, Ψ, χ_1). Side chain−side chain and side chain−backbone interaction potentials were parametrized to fit the potential mean forces of corresponding all-atom simulations. The force field has been applied in molecular dynamics simulations of several proteins of 56−108 AA residues whose X-ray crystal and/or NMR structures are available. Starting from the crystal structures, each protein was simulated for about 100 ns. The Cα RMSDs of the calculated structures are 2.4−4.2 Å with respect to the crystal and/or NMR structures, which are still larger than but close to those of all-atom simulations (1.1−3.6 Å). Starting from the PDB structure of malate synthase G of 723 AA residues, the wall-clock time of a 30 ns simulation is about three days on a 2.65 GHz dual-core CPU. The RMSD to the experimental structure is about 4.3 Å. These results implicate the applicability of the force field in the study of protein structures.

Introduction

Protein modeling with molecular mechanics (MM) force fields plays an important role in computational biology.[1−3] However, handling real systems with the current popular all-atom force fields is limited in both size and time scale due to the computationally demanding nature of these force fields.[4−6] In recent years, there have been increasing efforts in developing coarse-grained (CG) force fields. Multiple interaction sites of biomolecules are simplified to one site, leading to a great increase in simulation capability. These force fields describe proteins or DNA/RNAs at various resolutions, ranging from the residue-based level,[7−22] to the intermediate level,[23−34] and to the united-atom-based level.[35−41] They have extended greatly the temporal and spatial scale of simulations.[42−45] Some of them already have predictive capabilities.[24,26,31,33,38,39]

Coping with environments such as explicit water requires a great deal of calculation time. Implicit solvent models are one way of reducing the computational cost. The generalized Born (GB) solvent model has been carefully optimized so that it fits well with the current all-atom force fields.[1,46] Most CG force fields also do not consider solvent explicitly but treat solvent-induced effects such as hydrophobic interactions as pairwise additive interactions.[7−13,20,23−34,38,39] The implicit solvent model may not be that efficient when the system becomes more complex, such as in the case of a heterogeneous protein/water/membrane system, in which the details of the membrane used are often of great interest. An alternative approach is to treat the environment explicitly in

* Corresponding author e-mail: chydwu@ust.hk.
[†] The Hong Kong University of Science & Technology.
[‡] Peking University Shenzhen Graduate School.
[§] Peking University.

10.1021/ct1003127
Published on Web 10/25/2010

a CG manner.[14−19,21,22,35,37,40,41] A good example is the MARTINI force field derived by Marrink and co-workers,[15−17] in which roughly four heavy atoms are reduced into one site for both the protein and environment. This force field has been successfully applied to protein/membrane interactions and membrane dynamics with a tremendous increase in efficacy.[18,19] However, due to the loss of atomistic information of proteins in their model, the restraints on native contacts are needed to maintain the native structures of proteins.

Recently, we have been pursuing a protein force field[40,41] that can be coupled with Marrink's CG environment but possesses adequate atomistic details for protein folding. By simulating the helix−coil transition of polyalanine-based peptides, we have initially demonstrated that such a force field coupled with the CG solvent is indeed possible.[40] Encouraged by the results, we are concentrating our effort on the extension of this force field to the simulation of real proteins.

To properly represent a protein, it is necessary to correctly describe the backbone (ϕ, Ψ) conformational preference of all 20 amino acids.[47−52] For the amino acids with rotating side chains, the degree of freedom (χ) of the side chains needs to be considered.[53] Fitting gas phase Ramachandran plots of alanine and glycine dipeptides with high-level quantum mechanics (QM) calculations was a conventional way of optimizing backbone potentials (ϕ, Ψ).[54−60] But during the development of the CHARMM−CMAP force field,[49] Brooks and co-workers found that even if the force field can reproduce exactly the high-level QM backbone potential in a vacuum, the systematic deviation in the (ϕ, Ψ) of the α helix and β sheet from experimental data is evident. They pointed out that the MM force field for the gas phase may not capture well the solvent dependence of the (ϕ, Ψ) potential in reality. Duan and co-workers[47] reoptimized the backbone potentials of the AMBER force field according to the high-level QM potential surface in a medium of $\varepsilon = 4.0$. The improved force field can reproduce well the experimental structural properties of both dipeptides and short polypeptides.

More recently, Liu and co-workers[50] remodified the backbone potential of the GROMOS force field with a QM potential surface in water ($\varepsilon = 78.0$) as a reference. Moreover, they empirically adjusted the parameters by fitting the free energy surface, instead of the potential surface, of their model to the QM data. The optimized force field is significantly improved in its ability to reproduce the native structures of proteins in time-intensive simulations. We here chose a similar way to optimize the backbone parameters of our force field through the calculation of free energy surfaces. The statistical Ramachandran surfaces from PDB were used as references.

In addition to the backbone parameters, we also need to optimize the parameters of the nonbonded interactions among atoms in 20 amino acids. Following a thermodynamics-based approach,[14−17] a large part of these parameters were obtained by reproducing the experimental thermodynamic properties of organic molecules such as self-solvation free energies and hydration free energies.[41] However, the parameters among polar/charged groups remain absent due to a severe lack of direct experimental data on the pairwise interactions between polar/charged groups in solvent environments. As such, we plan to resort to explicit water simulations, from which the potential mean force (PMF) between polar/charged groups can be derived for the fitting. There are a number of successful precedent applications of explicit water PMF to the developments of force fields such as MARTINI[61] and UNRES[62] force fields. A similar strategy was used by Chen et al. for a GB implicit solvent model.[63]

In this paper, we report our recent work toward the full parametrization of our force field, namely, the Protein in Atomistic details coupled with Coarse-grained Environment (PACE). Specifically, the backbone (ϕ, Ψ) potential was optimized to reproduce the Ramachandran plots from a protein data bank (PDB) coil library through aqueous simulations of dipeptides of 20 amino acids. The conformational (χ) distributions of side chains and the backbone preferences in different side chain conformations were also used in the fitting. The interactions between polar and charged sites in proteins were optimized by fitting the PMFs of simulations with the OPLS-AA/L force field in explicit water. Together with the previous parameters, our force field is already able to tackle proteins with real sequences. As a preliminary test, tens to hundreds of nanoseconds of aqueous simulations of several proteins (∼50 AA to ∼700 AA) were performed. Native structures were well preserved in all of the simulations (RMSD < 0.43 nm).

Models and Methods

The Coarse-Grained Protein Model. Our protein model is united-atom-based (UA). As shown in Figure 1, normally, each heavy atom together with the attached hydrogen atoms is represented by one site. But the hydrogen atoms in backbone amide groups and in the side chains of Asn, Gln, Trp, and His are also explicitly represented to better account for their hydrogen-bonding property. The total energy of our model is expressed in eq 1:

$$E = E_{\text{angle}} + E_{\text{torsion}} + E_{\text{14pair}} + E_{\text{improper}} + E_{\text{nonbond}} + E_{\text{HB}} \quad (1)$$

The detailed description of the above terms and their parametrization can be found in our previous papers.[40,41] The first four terms describe the bonded interactions between the sites that are connected through not more than three covalent bonds. All bond lengths are constrained at their equilibrium values. All bond angles are restrained with a harmonic potential E_{angle} at their equilibrium values with a force constant of $K = 300$ kJ/mol/rad^2. The equilibrium values of bond lengths and angles are obtained through the QM calculations of 15 small molecules. All the values for bond length and bend angle are illustrated in Figure S1 in the Supporting Information (SI). E_{torsion} and E_{14pair} describe the potential energy of a dihedral angle about a rotating bond, which has forms in eqs 2 and 3. E_{14pair} stands for the interaction between sites separated by three bonds. For an amino acid, its side chain torsion potential is optimized by fitting the QM torsional potential

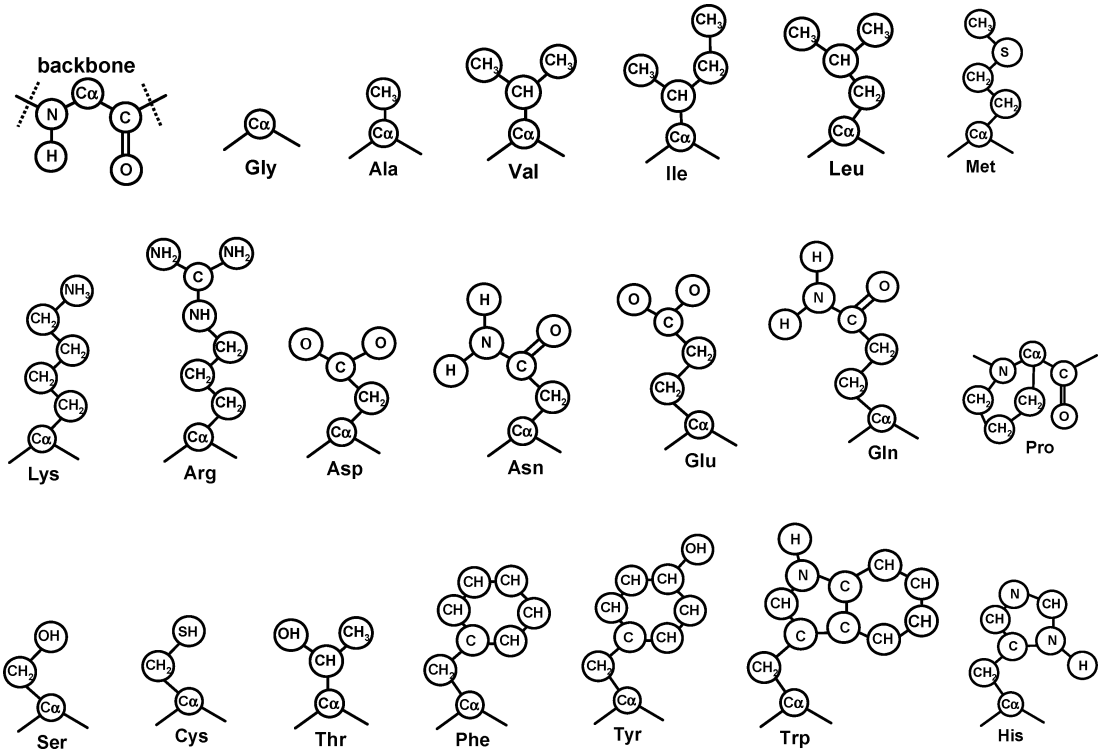

Figure 1. Schematic representations of amino acids in our model. Each circle indicates a site that is explicitly represented.

of 24 minima and 22 rotation barriers of simple molecules. But for the backbone part (ϕ, Ψ), the parameters will be optimized by fitting the experimental (ϕ, Ψ) maps in this work. E_{improper} is used to keep planar geometries and chiral centers in molecules. It imposes a harmonic potential on a dihedral of four sites with $K = 300$ kJ/mol/rad^2. All of the parameters for bonded terms can be found in Tables S1−S2 in the SI.

$$E_{\text{torsion}} = \sum_i K_{\text{torsion},i}[1 + \cos(n_i\zeta_i - \zeta_{0,i})] \qquad (2)$$

$$E_{\text{14pair}} = \sum_{1-4\text{relationship}} 4\varepsilon_{14-ij}\left(\frac{\delta_{14-ij}^{12}}{r^{12}} - \frac{\delta_{14-ij}^{6}}{r^{6}}\right) \qquad (3)$$

E_{nonbond} and E_{HB} describe the interactions between sites beyond three-bond connections. E_{nonbond} is used for isotropic nonbonded interactions. These interactions consist of both van der Waals interactions and electrostatic interactions that are normally treated separately in all-atom force fields.[58−60] But for simplicity, in our model, only a Lennard-Jones potential (eq 4) is used to represent the overall effect. Thus, all of the atomistic charges are set to zero in our current force field.

$$E_{\text{nonbond}} = \sum_{i\neq j} 4\varepsilon_{ij}\left(\frac{\delta_{ij}^{12}}{r^{12}} - \frac{\delta_{ij}^{6}}{r^{6}}\right) \qquad (4)$$

The CG water developed by Marrink et al. is a vdW sphere, representing a cluster of four water molecules. The LJ parameters ε and δ of the CG water site are 5.0 kJ/mol and 0.47 nm, respectively, which are optimized through reproducing the density and compressibility of pure water. We have treated the interaction between CG water and protein sites with eq 4, with δ_{ij} as the summation of vdW radii of CG water and protein sites while ε_{ij}'s were optimized by fitting hydration free energies of 35 organic compounds.[41] The average deviation from experimental data is about 1.1 kJ/mol. In our previous study, self-solvation free energies (or free energies of evaporation) of eight organic compounds were used to optimize E_{nonbond} parameters (ε_{ij} and δ_{ij}) for the interactions between protein sites.[41] These compounds include cyclohexane, *n*-pentane, isopentane, benzene, diethyl ether, 2,3-dimethyl-2-butene, triethylamine, and dimethylsulfide. The average errors for density and self-solvation free energy are about 3.2% and 0.7 kJ/mol, respectively. These parameters of nonbonded parts are listed in Tables S3−S4 in the SI. It should be noted that to reproduce the solvation free energies and densities of these eight compounds only requires that E_{nonbond} parameters for the interactions between nonpolar sites and between nonpolar and polar sites are optimized. Thus, the interactions between polar sites such as HBs were left unparametrized.[41] Therefore, one of our main focuses in this work is to obtain the parameters for the interactions between polar sites.

E_{HB} describes hydrogen bond (HB) interactions between two groups of sites such as amide groups. Because HB interactions have directionality, we have devised a set of attractive and repulsive potentials[40] which are simultaneously employed to maintain the directionality of HB between backbone amide groups (Figure 2a and eq 5), which has been adopted in the study by Dokholyan and co-workers.[39]

$$E_{HB} = \sum_{|i-j|>2} [4\varepsilon_{attr}\left(\frac{\delta_{Oi-NHj}^{12}}{r_{Oi-NHj}^{12}} - \frac{\delta_{Oi-NHj}^{6}}{r_{Oi-NHj}^{6}}\right) + 4\varepsilon_{rep}\frac{\delta_{Oi-C\alpha j}^{12}}{r_{Oi-C\alpha j}^{12}} + 4\varepsilon_{rep}\frac{\delta_{Oi-Cj-1}^{12}}{r_{Oi-Cj-1}^{12}} + 4\varepsilon_{rep}\frac{\delta_{Ci-NHj}^{12}}{r_{Cj-NHj}^{12}}] \quad (5)$$

Here, only the HB between two amides that are separated by at least two residues is considered. All of the parameters have been optimized by fitting the change of free energy, entropy, and enthalpy of the helix−coil simulations of polyalanine-based peptides in our previous work.[40] All of the parameters are kept unchanged in this work except for ε_{attr} between N and O atoms, which will be further optimized in this work. The parameters for E_{HB} are shown in Table S5 in the SI.

The backbone HB potential is not applicable to the HB involving side chain amide groups of Asn and Gln. This is because in these amide groups an amide nitrogen atom can have two hydrogen atoms as donors, but the original HB potential can only handle one hydrogen donor on each nitrogen atom. Therefore, these hydrogen atoms need to be explicitly represented, and another type of HB potential should be used (Figure 2b and eq 6).

$$E'_{HB} = \sum_{|i-j|>2}\left[4\varepsilon'_{attr}\left(\frac{\delta'^{12}_{O-H}}{r_{O-H}^{12}} - \frac{\delta'^{6}_{O-H}}{r_{O-H}^{6}}\right) + 4\varepsilon'_{rep}\frac{\delta'^{12}_{C-H}}{r_{C-H}^{12}} + 4\varepsilon'_{rep}\frac{\delta'^{12}_{O-N}}{r_{O-N}^{12}}\right] \quad (6)$$

We have applied this approach to the HB interactions between the side chains of Asp, Asn, Glu, Gln, His, and Trp (Figure 1) and between these side chains and the backbone amide groups. The optimization of parameters for backbone−backbone interactions is based on the contents of secondary structures of peptides. The parameters for side chain−side chain and backbone−side chain HB interactions are optimized by fitting the all-atom PMF. The details of the optimization will be discussed in the Results and Discussions.

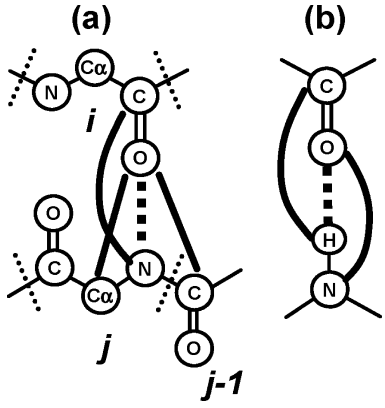

Figure 2. Schematic representation of hydrogen bond potentials (a) between backbone amide groups and (b) between donors and acceptors in side chains. The solid lines denote repulsive potentials, and the dotted lines denote attractive potentials.

The complete force field package is available upon readers' requests. This package includes the force field files that are compatible with the GROMACS software (v3.3) and a written code that can automatically modify topology files so that our force field can be used with GROMACS.

Dipeptide Simulations. All of the simulations were performed with the GROMACS software packages (version 3.3.1).[64] A dipeptide molecule (Ace−Xxx−NMe) that is capped at the two ends was placed in a dodecahedron box so that the minimum distance between box edges and the molecule was about 1.4 nm. The box was filled with 400−500 CG water particles. All bonds are constrained with the LINCS algorithm.[65] All nonbonded interactions were shifted to zero between distances of 0.9 and 1.2 nm. The system was optimized with the steepest descent methods. After a 5000-step optimization, the system was pre-equilibrated at 300 K and 1.0 atm for 50 ps. The pressure and the temperature were controlled by a thermostat and a pressure bath with coupling constants of 0.1 and 0.5 ps, respectively.[66] After the pre-equilibrium, the system was heated at 700 K with a constant volume for 10 ns. The last 5 ns of the heating simulation were used to generate the starting conformations for production simulations. Dummy atoms were used to remove the fastest motion of the system involved with hydrogen.[67] These dummy atoms are virtual interaction sites whose positions are determined by three nearby heavy atoms. Thus, in all of the normal molecular dynamics (MD) simulations including protein simulations, a large time step of 6 fs can be used, which has been shown to satisfy energy conservation during simulations.[67] As in our force field, all of the parameters related to hydrogen are optimized with the treatment of dummy sites; we suggest that for consistency, dummy sites should be used for explicit hydrogen atoms in simulations with our force field.

Replica exchange molecular dynamics (REMD) simulations provide an efficient method to perform equilibrium simulations.[68,69] After combining with our fast UA model, the REMD method becomes beneficial to parametrization. Our REMD simulations contained 16 replicas with temperatures ranging from 300 to 431 K at 1 atm. Each replica started with a different conformation generated from the heating simulation at 700 K. Exchanges were attempted every 2 ps. The successful exchange rate was about ∼15%. For each REMD simulation, each replica lasted for 50 ns. The last 40 ns of the simulation at 300 K were used for analysis.

Finally, in all of the REMD simulations of dipeptides, the Newton equation was integrated every 10 fs. Since most sites in our model represent only a single atom and are connected directly by strong covalent bonds, the system is prone to crashes at large time steps. This is particularly obvious in the REMD simulations, as simulation at a high temperature (up to 431 K) is involved. To avoid crashes, we treated all explicit hydrogen atoms as dummy sites and tripled the mass of all proteins. Our previous studies and the peptide simulations in an accompanying paper show that the mass scaling plus the use of dummy atoms should not affect the thermodynamic properties of the system, which is our main interest in the current study.[40]

PMF Calculations. PMF calculations were performed in a rectangular box with a size of 7.0 × 2.9 × 2.9 nm. A pair of small molecules was put in the center of the box solvated by about 500 CG water particles. To obtain the PMF of solutes at a particular orientation, solutes were constrained to move along a straight line. The PMF was calculated using the free energy perturbation method in this work. Free energy perturbation calculates the free energy change of a solute pair separated at different distances. The positions of the solute pair are restrained so that the pair is either in a close contact (∼0.26 nm) or widely separated (∼2.2 nm). The positions can be made a function of a coupling parameter (λ). As λ slowly varies from zero to unity (10 000 000 perturbation steps with a time step of 6 fs), the solutes are moved away from each other. The free energy difference between the pair in any intermediate state and the close contact state is determined as the accumulated energy difference between the conformation corresponding to λ and that corresponding to $\lambda + \delta\lambda$ at each perturbation step. The free energy at the longest distance is set to zero as the reference point. In each perturbation simulation, the free energies of 100 intermediate states were chosen for PMF plots. For each PMF calculation, six to eight perturbation simulations were performed to generate an averaged PMF curve. All simulations were kept at 300 K and 1 atm throughout the whole simulation.

PMFs were also calculated in all-atom simulations for comparison when necessary. The calculation protocol is similar to that of the CG simulation. The OPLS-AA force field[70] and TIP3P water model were used. As long-range electrostatics might be important in PMF calculations, PME summation[71] was used in all atomistic simulations. In the CG PMF calculations, the shift potential was used for the nonbonded interactions like our normal MD and REMD simulations. About 2000 water molecules were placed in a box. Each perturbation simulation was performed for 5 000 000 perturbation steps with a time step of 2 fs. A total of 100 intermediate states were chosen in each simulation, and six to eight perturbation simulations were carried out for each PMF calculation.

Statistical Analysis. Protein X-ray crystal structures of a resolution < 0.2 nm andwith an R factor < 0.2 were selected and downloaded from the Protein Data Bank (PDB),[72] with a 50% sequence identity cutoff. There are a total of 4220 protein structures and 2.0×10^6 residues. When there are n identical chains in a protein structure, each chain is counted with a $1/n$ statistical weight. To obtain the local conformational preferences, we only chose the coil residues that are not adjacent to any secondary structures,[73] as previously suggested.[74] The residues that have backbone atoms with temperature factor $B > 36$ or preceding prolines[75] are also excluded from our new restricted coil library (Coil-R for short). Furthermore, we exclude the residues preceding a turn-like α conformation ($-60° < \phi < +60°$, both α_L and α_R). The reason is because the side chains of some amino acids such as Ser and Asx can form hydrogen bonds with the backbones of the next residues if they are in the α conformation.

Secondary Structure Analysis. As suggested by Garcia et al.,[76] a residue is considered to be helical only if this residue and both of its neighboring residues have their backbone dihedral (ϕ, Ψ) within ($-60° \pm 30°$, $-47° \pm 30°$). If the backbone dihedral (ϕ, Ψ) of two residues are within ($-135° \pm 45°$, $+135° \pm 45°$) and there is at least a HB between the adjacent backbone amide groups to the two residues, the two residues are considered as having β sheets. A HB is considered as formed when the donor−acceptor distance is shorter than 0.35 nm and the donor−hydrogen−acceptor angle is larger than 120.0°.

Results and Discussions

Parameterization of Backbone Potentials. As a first step, we carried out parametrization of backbone potentials using the dipeptides of glycine and alanine. We used our recently reported statistical potentials derived from a coil library of the high-resolution X-ray crystal structure database.[73] Wang and co-workers have shown that the high-level QM potential surfaces of alanine and glycine dipeptides in an implicit water solvent are remarkably similar to the statistical results.[48] A similar observation was reported by Hu et al.[52] in their QM/MM simulations of alanine and glycine dipeptides. Table 1 shows a comparison between the surfaces from our coil library and the surfaces from the high-level QM calculation in a water solvent. The two types of surfaces agree well with each other in not only the depths of minima but also the heights of transition barriers. This is expected for the coil library, which reflects the conformational features of amino acids in solvent environments.

Our coil library[73] purposely avoids the nonlocal effects of other amino acids such as vdW and HB interactions. Therefore, the derived statistical surfaces correspond to the intrinsic conformational preference of amino acids, which is supported by the comparison with the QM results (Table 1). In folded proteins, besides the intrinsic preference of amino acids, nonlocal vdW and HB interactions are also important. These nonlocal effects ($E_{nonbond}$ and E_{HB}) are separately treated and parametrized in our force field. This strategy of combining the intrinsic preference and the nonlocal effects appears to work well, as supported by our protein simulations in this work (see later) and the folding simulations of peptides in our accompanying paper.

The optimization of the backbone dihedral (ϕ, Ψ) potential is empirical; that is, each further adjustment of parameters is based on the free energy surface of the last simulation. All of the parameters of ϕ and Ψ are optimized, including the dihedral potentials of ϕ and Ψ and all of the related 1−4 pair interactions. The optimized values from our previous study[40] are used as the starting input. We also found that several nonbonded interactions (Figure 3) are important to the (ϕ, Ψ) free energy surface, such as $N_i \cdots H_{i+1}$, $H_i \cdots H_{i+1}$, $C_{\beta i} \cdots H_{i+1}$, $O_{i-1} \cdots O_i$, $O_{i-1} \cdots C_{\beta i}$, $O_{i-1} \cdots N_{i+1}$, and $O_{i-1} \cdots C_i$. Some of these interactions, such as $O_{i-1} \cdots O_i$, $O_{i-1} \cdots N_{i+1}$, and $O_{i-1} \cdots C_{\beta i}$, have been suggested to be the most effective in shaping the (ϕ, Ψ) surface in a geometric analysis of the PDB structures by Ho et al.[77] Therefore, the interaction parameters of these atom pairs are also optimized to reproduce the target surface.

Table 1. Comparison of Positions (ϕ, Ψ in deg); Free Energy Difference (kJ/mol) of β, PPII, PPII′, α_R, and α_L Conformations; and Barrier Heights (kJ/mol) between the Conformations for Different Methods Including the Coil Library, Quantum Mechanics Calculations by Wang et al.,[48] and our UA Model[a,b]

	β	PPII	PPII′	α_R	α_L	$\beta\rightarrow\alpha_R$	$\beta\rightarrow$PPII	PPII$\rightarrow\alpha_L$
				Ala				
coil	(−155,155)	(−65,145)	(55, −135)	(−65, −35)	(55,45)	8	4	18
	3.0	0.0	12.4	2.1	6.9			
QM	(−156,144)	(−64,142)	N/A	(−70, −32)	(59,41)	7	1	24
	0.0	0.0		1.0	4.0			
UA	(−165,145)	(−75,145)	(65, −145)	(−75, −25)	(65,35)	8	4	20
	2.4 ± 0.1	0.0	10.4 ± 2.2	1.4 ± 0.2	7.7 ± 1.1			
				Gly				
coil	(−175,175)	(−75,165)	(75, −165)	(−85, −15)	(75,15)	9	4	20
	3.3	1.4	1.2	3.0	0.0			
QM	(−166,173)	(−66,153)	(66, −153)	(−66, −26)	(66,26)	7		
	3.3	1.2	1.2	0.0	0.0			
UA	(−175,175)	(−75,145)	(75, −145)	(75,25)	(75, 25)	8	6	16
	3.7 ± 0.2	1.5 ± 0.4	1.7 ± 0.4	0.1 ± 0.5	0.0			

[a] The calculation is made at the MP2/cc-pVTZ//HF/6-31G** level in ε = 78.0 medium by Wang et al.[48] [b] The REMD simulations of dipeptide in CG water are performed at 300 K. Standard deviations were estimated from block averages with a block size of 10 ns.

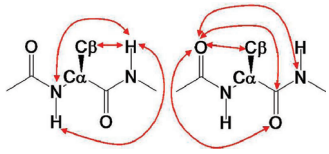

Figure 3. Schematic representation of remodified intraresidue nonbonded interactions, as denoted by red arrows.

Table 2. Summary of the Parameters of n, ζ_0 (deg), and $K_{torsion}$ (kJ/mol) in eq 2 for ϕ ($\angle$C−N−C_α−C) and Ψ ($\angle$N−C_α−C−N); the Parameters of ε_{14} (kJ/mol) and δ_{14} (nm) for Related 1−4 Pairs; and the Parameters of ε (kJ/mol) and δ (nm) for Remodified Intrabackbone Lennard-Jones Interactions

AA	dihedral	$K_{torsion}$	n	ζ_0
Ala	ϕ	2.5	3	0
		1	1	−180
	Ψ	4	2	−180
		1	1	0
Gly	ϕ	2.5	3	0
		3.5	1	−180
	Ψ	4	2	−180
		1	1	−180

1−4 pair	ε_{14}	δ_{14}	1−4 pair	ε_{14}	δ_{14}
C−C	0.9	0.23	C−C_β	0.9	0.28
H−C_β	0	0	H−C	0	0
N−O	0.9	0.275	N−N	0.9	0.24
C_β−O	0.9	0.28	C_β−N	0.9	0.31

nonbond	ε	δ	nonbond	ε	δ
O_{i-1}−C_i[a]	0.6	0.27	O_{i-1}−N_{i+1}	0.894	0.26
N_i−H_{i+1}	2.98	0.183	O_{i-1}−$C_{\beta i}$	0.894	0.32
O_{i-1}−O_i	0.8	0.33	H_i−H_{i+1}[b]	1.95	0.22
$C_{\beta i}$−H_{i+1}	0.447	0.315			

[a] The subscripts indicate the identities of atoms in amino acids (refer to Figure 3). [b] For these interactions, only the repulsive part of the Lennard-Jones terms, $4\varepsilon\delta^{12}/r^{12}$, is used.

Results for Alanine Dipeptide. The optimized parameters for alanine dipeptides are listed in Table 2. The simulated (ϕ, Ψ) map with the optimized parameters is shown in Figure 4c. Compared to the (ϕ, Ψ) map from the coil library (Figure 4a), all of the major minima including α_R, α_L, β, and PPII are reproduced by our UA model. Table 1 contains a detailed comparison of the positions and relative free energies of these minima and the transition barrier heights between the minima from the coil library, the high-level QM calculations, and our simulations. The global minimum of the (ϕ, Ψ) surface in our simulations is the PPII conformer for Ala, in agreement with both the statistical analysis and QM calculations. The relative free energies of β, α_R, and α_L to PPII are 2.4, 1.4, and 7.7 kJ/mol, respectively. Our result of the stability of the β-conformer is closer to the statistical value (3.0 kJ/mol) than the QM value (0.0 kJ/mol). The relative stability (1.4 kJ/mol) of α_R conformers in our simulations is however closer to the QM results (1.0 kJ/mol) than to the statistical ones (2.1 kJ/mol). As for the α_L conformers, their stability estimated with our model is the same as that of the statistical

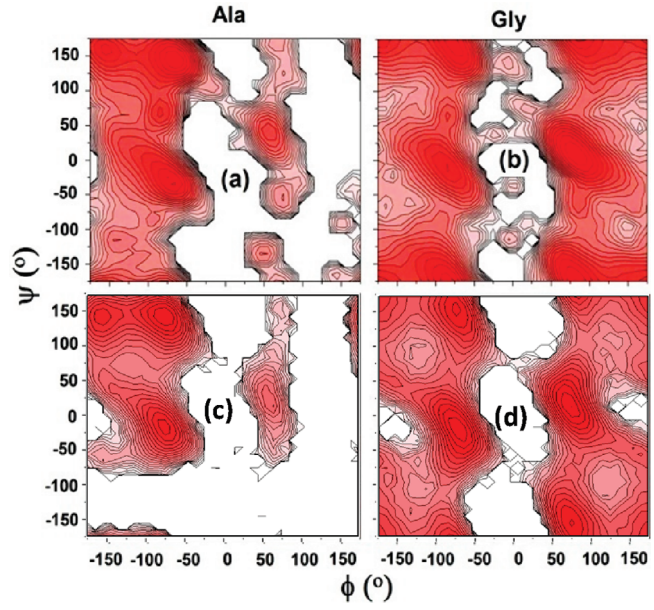

Figure 4. Statistical (ϕ, Ψ) PMFs of (a) Ala and (b) Gly dipeptides from the coil library. The calculated (ϕ, Ψ) PMFs with our UA model are shown in c for Ala and d for Gly dipeptides. The gap between the contour lines indicates a free energy difference of 1 kJ/mol.

analysis. Considering the previously examined force fields with the relative stability of α_L at more than 12.5 kJ/mol,[52] our model has considerably raised the stability of α_L. Besides the local minima, the barriers between the minima are also well reproduced. The heights of the barriers between β and α_R, between β and PPII, and between PPII and α_L are 8, 4, and 20 kJ/mol, respectively, which are very close to the statistical values (8, 4, and 18 kJ/mol, respectively).

Results for Glycine Dipeptide. As shown in Figure 4, Ala and Gly have quite different Ramachandran plots, due to the replacement of C_β in Ala by H_α in Gly. As our force field implicitly represents most H atoms, including H_α, the difference between Ala and Gly has to be reflected in another way, such as using different backbone potentials for Ala and Gly. We found that most of the backbone parameters remain unchanged, except for a minor change in the torsional potential of ϕ (Table 2), which is already good enough to reproduce the statistical result of Gly.

Figure 4d shows our (ϕ, Ψ) free energy surface of Gly. Since Gly has no chiral center, its (ϕ, Ψ) map should have C_2 symmetry about ($\phi \sim 0$, $\Psi \sim 0$). In the plot of Gly from the coil library (Figure 4b), an asymmetry appears. That is, α_L is more populated than α_R. This indicates a trace of the coupling between Gly and its adjacent residues in proteins despite our effort to remove the coupling (see Models and Methods). Except for this, the two plots (Figure 4b and d) are quite similar. Five major minima are well reproduced, including α_R/α_L ($\phi \sim \pm\pi/2$, $\Psi \sim 0$), PPII/PPII′ ($\phi \sim \pm\pi/2$, $\Psi \sim \pi$) and β ($\phi \sim \pi$, $\Psi \sim \pi$). The global minima are α_L and α_R, and the relative free energies of PPII, PPII′, and β are 1.5, 1.7, and 3.7 kJ/mol, respectively, compared quite favorably with the statistical results (1.4, 1.2, and 3.3 kJ/mol, respectively) as well as the QM-calculated results (1.2, 1.2, and 3.3 kJ/mol).

Parameterization of Side Chain Potentials (χ). Except for Ala, Gly, and Pro, other amino acids have rotating side chains. In particular, statistical analyses[73,78] have revealed that the degree of freedom of a side chain dihedral angle χ ($\angle$N−Cα−Cβ−Cγ) plays a critical role in determining the conformational features of amino acids. Following the convention, according to the range of the χ angle, side chains are defined as being in the g+ ($-120° < \chi < 0°$), g− ($0° < \chi < 120°$), and t ($120° < \chi < 240°$) states (IUPAC-IUB Commission on Biochemical Nomenclature, 1970).

Just like the coupling between ϕ and Ψ, the coupling between χ and (ϕ, Ψ) is also controlled by a set of nonbonded interactions between special atom pairs.[78] Figure 5 illustrates these interactions. They can be divided into two kinds. The first kind is the steric repulsion between side chains (atoms at γ and δ positions) and their neighboring backbone amides (C, N, O, and H). In our testing simulations, normal nonbonded parameters appeared too repulsive for these interactions. The modified parameters should have reduced repulsive forces. One exception is the repulsion involving backbone H sites. Our previous QM calculations have shown that the repulsion between C_γ and backbone H sites is important in determining the relative energies of α_R and β conformations for amino acids such as Val.[40] Therefore, this repulsion was taken into account. The second

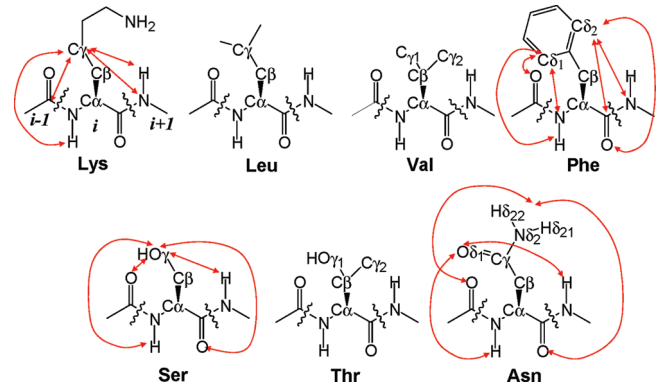

Figure 5. Schematic representation of local side chain−backbone interactions that are optimized.

kind of remodified interactions include those polar interactions between side chain polar groups and backbone amides. The normal nonbonded parameters can be too strong for these interactions. As revealed by our testing simulations, these parameters may distort the conformational distributions of backbones and side chains from the statistical results. This is especially severe for amino acids such as Asp, Asn, Ser, Thr, and His, all of which have polar groups close to their backbones. The local side chain−backbone interactions are major parameters to be optimized.

In an ideal case, all amino acids would share exactly the same set of parameters, which requires a global optimization. Although such global optimization will be expensive with our empirical fitting procedures, we are heading toward this end. The current approach is a compromise. That is, all backbone parameters including most of the ϕ, Ψ, and χ torsional terms and the intraresidue interaction terms are the same for all amino acids. However, the torsional terms of the ϕ, Ψ, and χ dihedral angles, the multiplicities (n) of which are one, were fine-tuned for individual amino acids in order to account for the distinct α_R, β, or even α_L propensities of different amino acids. Although this leads to a loss of transferability of the force field, it significantly simplifies the parametrization task and does not appear to lose applicability to real peptide systems. In the accompanying paper, we demonstrate that starting from random coils our force field can fold peptides with different sequences into their native structures.

Our parametrizations of the local interactions and torsional potentials are guided by two principals: (1) to match the populations of side chain rotamers (g+, g−, and t) and (2) to fit the Ramachandran plots of the three rotamers of each amino acid. In each side chain rotamer, there are four important minima including α_L, α_R, PPII, and β. For each amino acid, there are up to 12 minima (or basin). Their positions and basin depths can be fitted. Thus, the reference data should be more than enough to optimize all of the parameters. It should be noted that for most of the amino acids, their most favorable side chain conformations make up more than 50% of the population. Therefore, during the optimization, the backbone conformational distribution in the most favorable side chain conformations has a higher priority to be fitted than those in the other side chain conformations. All of the optimized parameters are summarized in Table 3.

Table 3. Summary of the Parameters of n, ζ_0 (deg), and $K_{torsion}$ (kJ/mol) in eq 2 for ϕ ($\angle C-N-C_\alpha-C$), Ψ ($\angle N-C_\alpha-C-N$), and c ($\angle N-C_\alpha-C_\beta-\gamma$); the 1−4 Pair Parameters of ε_{14} (kJ/mol) and δ_{14} (nm) about χ; and the Parameters of ε (kJ/mol) and δ (nm) for Remodified Local Backbone−Side Chain Lennard-Jones Interactions

	dihedral	$K_{torsion}$	n	ζ_0	1−4 pair	ε_{14}	δ_{14}	non-bond	ε	δ
All	ϕ	2.5	3	0						
	Ψ	2	4	−180						
	χ	4.9	3	0						
Lys/Arg	χ	0.5	1	180	$C_{\gamma i}-N_i$[a]	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$	0.447	0.315
Gln/Glu	Ψ (Leu)	0.75	1	0	$C_{\gamma 1}-C_i$	0.1	0.33	$C_{\gamma i}-N_{i+1}$	0.447	0.35
Met/Leu	χ (Leu)	0.75	1	120				$C_{\gamma i}-C_{i-1}$	0.224	0.33
Val	ϕ	2	1	0	$C_{\gamma 1,i}/C_{\gamma 2,i}-N_i$	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$	0.447	0.29
Ile	Ψ	2	1	0	$C_{\gamma 1,i}/C_{\gamma 2,i}-C_i$	0.1	0.33	$C_{\gamma i}-N_{i+1}$	0.447	0.35
								$C_{\gamma i}-C_{i-1}$	0.224	0.33
Phe	ϕ	1.75	1	0	$C_{\gamma i}-N_i$	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$	0.224	0.3
Tyr	Ψ	2.5	1	0	$C_{\gamma 1}-C_i$	0.1	0.3	$C_{\gamma i}-C_{i-1}$	0.224	0.3
Trp	χ	0.4	1	120				$C_{\gamma i}-N_{i+1}$	0.447	0.32
								$C_{\delta 1,i}/C_{\delta 2,i}-C_{i-1}/C_i$	0.224	0.3
								$C_{\delta 1,i}/C_{\delta 2,i}-N_i$	0.447	0.3
								$C_{\delta 1,i}/C_{\delta 2,i}-N_{i+1}$	0.447	0.34
								$C_{\delta 1,i}/C_{\delta 2,i}-H_i$	0.224	0.3
								$C_{\delta 1,i}/C_{\delta 2,i}-O_{i-1}$	0.447	0.29
Asn	ϕ	1	1	0	$C_{\gamma i}-N_i$	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$[b]	7.5	0.235
	ϕ	2.75	1	−120	$C_{\gamma i}-C_i$	0.1	0.3	$C_{\gamma i}-C_{i-1}$	0.224	0.3
	χ	2	1	−150				$C_{\gamma i}-N_{i+1}$	0.447	0.32
								$N_{\delta 2,i}-N_i$	0.447	0.3
								$N_{\delta 2,i}-C_{i-1}/C_i$	0.224	0.3
								$N_{\delta 2,i}-N_{i+1}$	0.447	0.34
								$H_{\delta 21,i}-O_i$[c]	22	0.16
								$O_{\delta 1,i}-H_i/H_{i+1}$[c]	25	0.16
								$O_{\delta 1,i}-O_{i-1}/O_i$[b]	15	0.235
								$O_{\delta 1,i}-C_{i-1}/C_i$	0.447	0.29
Asp	ϕ	1.5	1	0	$C_{\gamma i}-N_i$	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$[b]	7.5	0.235
	ϕ	2	1	−120	$C_{\gamma i}-C_i$	0.1	0.3	$C_{\gamma i}-N_{i+1}$	0.447	0.32
	χ	2.75	1	165				$C_{\gamma i}-C_{i-1}$	0.25	0.36
								$O_{\delta 1,i}/O_{\delta 2,i}-O_{i-1}/O_i$[b]	15	0.235
								$O_{\delta 1,i}/O_{\delta 2,i}-C_{i-1}/C_i$	0.447	0.29
								$O_{\delta 1,i}/O_{\delta 2,i}-H_i$[c]	25	0.16
								$O_{\delta 1,i}/O_{\delta 2,i}-H_{i+1}$[c]	19.2	0.162
								$C_{\gamma i}-O_i$	6	0.26
Cys	Ψ	0.5	1	0	$S_{\gamma i}-N_i$	0.1	0.31	$S_{\gamma i}-N_{i+1}$	0.447	0.35
	χ	1	1	−120	$S_{\gamma i}-C_i$	0.1	0.35	$S_{\gamma i}-C_{i-1}$	0.224	0.33
Ser	ϕ	1.5	1	−120	$O_{\gamma i}-N_i$	0.1	0.3	$O_{\gamma i}-H_i/H_{i+1}$	2.5	0.22
	χ	1.5	1	−120	$O_{\gamma i}-C_i$	0.1	0.3	$O_{\gamma i}-N_{i+1}$	0.894	0.32
								$O_{\gamma i}-O_{i-1}$	2.5	0.27
								$O_{\gamma i}-O_i$	0.1	0.353
Thr	ϕ	2	1	0	$O_{\gamma 1,i}-N_i$	0.1	0.3	$C_{\gamma 2,i}-H_i/H_{i+1}$	0.447	0.29
	Ψ	2	1	0	$O_{\gamma 1,i}-C_i$	0.1	0.3	$C_{\gamma 2,i}-N_{i+1}$	0.447	0.35
	χ	1	1	−120	$C_{\gamma 2,i}-N_i$	0.1	0.29	$C_{\gamma 2,i}-C_{i-1}$	0.224	0.33
					$C_{\gamma 2,i}-C_i$	0.1	0.33	$O_{\gamma 1,i}-H_{i+1}$	2.5	0.22
								$O_{\gamma 1,i}-H_i$	0.1	0.288
								$O_{\gamma 1,i}-N_{i+1}$	0.894	0.32
								$O_{\gamma 1,i}-O_{i-1}$	2.5	0.27
								$O_{\gamma 1,i}-O_i$	0.1	0.353
Pro	Ψ	2.5	1	0	$C_{\gamma i}-N_i$	0.1	0.29	$C_{\gamma i}-H_i/H_{i+1}$	0.224	0.3
His	ϕ	2.25	1	0	$C_{\gamma i}-C_i$	0.1	0.3	$C_{\gamma i}-C_{i-1}$	0.224	0.3
	ϕ	2	1	−120				$C_{\gamma 1}-N_{i+1}$	0.447	0.32
	Ψ	1	1	0				$C_{\delta 2,i}-C_{i-1}/C_i$	0.224	0.3
	χ	0.3	1	120				$N_{\delta 1,i}-C_{i-1}/C_i$	0.224	0.3
								$C_{\delta 2,i}/N_{\delta 1,i}-H_i$	0.224	0.3
								$C_{\delta 2,i}/N_{\delta 1,i}-N_i$	0.447	0.3
								$C_{\delta 2,i}/N_{\delta 1,i}-N_{i+1}$	0.447	0.34
								$C_{\delta 2,i}-O_{i-1}$	0.447	0.29

[a] For these interactions, only a part of the Lennard-Jones terms, $4\varepsilon\delta^{12}/r^{12}$, is used to represent the repulsive forces between the pair of particles. [b] The subscripts indicate the identities of atoms in amino acids (refer to Figure 5). [c] The large values of ε are used for hydrogen bonding interactions. [d] A harmonic constraint at −63° with $K = 41$ kJ/mol rad^2 is applied to ϕ of Pro.

Results for Side Chain Conformers. Table 4 shows the preference of side chain conformers for all amino acids except for Ala, Gly, and Pro. Our results are obtained through REMD simulations of dipeptides in CG water with the optimized parameters that are listed in Tables 2 and 3. As shown in Figure 6, the populations of three side chain rotamers of all amino

Table 4. Population of Side Chain Rotamers (g−, g+, and t) of Various Amino Acids from Our UA Model[a] and the Coil Library

	g+		g−		t	
	CG	coil	CG	coil	CG	coil
Lys	0.712	0.649	0.057	0.106	0.231	0.246
Gln	0.712	0.640	0.070	0.114	0.218	0.246
Glu	0.723	0.601	0.061	0.139	0.216	0.260
Arg	0.701	0.606	0.069	0.166	0.230	0.228
Met	0.728	0.613	0.058	0.139	0.214	0.248
Leu	0.697	0.731	0.016	0.018	0.287	0.251
Val	0.613	0.574	0.290	0.333	0.097	0.093
Ile	0.511	0.612	0.362	0.267	0.127	0.121
Phe	0.629	0.516	0.074	0.190	0.296	0.295
Trp	0.526	0.447	0.119	0.233	0.355	0.320
Tyr	0.643	0.484	0.091	0.188	0.266	0.328
His	0.592	0.556	0.102	0.189	0.306	0.255
Cys	0.561	0.458	0.183	0.292	0.256	0.250
Ser	0.298	0.294	0.620	0.534	0.083	0.172
Thr	0.203	0.273	0.698	0.637	0.099	0.090
Asn	0.627	0.551	0.161	0.189	0.212	0.260
Asp	0.547	0.482	0.216	0.228	0.237	0.289
standard deviation	<0.020		<0.030		<0.020	

[a] All of the calculated populations are obtained through dipeptide simulations in solution at 300 K. Standard deviations are estimated from block averages with a block size of 10 ns.

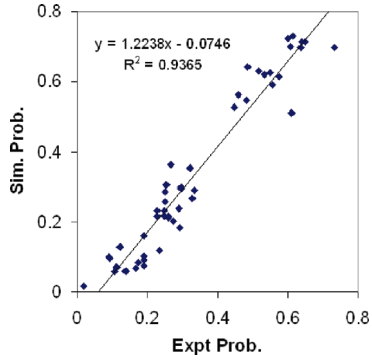

Figure 6. Plot of probabilities of side-chain rotamers (g+, g−, and t) of all amino acids except for Ala, Gly, and Pro in the coil library against the ones in our simulations.

acids expect for Ala, Gly, and Pro derived from our model match well with those from the coil library. The correlation coefficient R^2 is about 0.94, and the slope of the fitting line is 1.22. Interestingly, we found that simulations with the OPLS-AA/L and AMBERff03 force fields give no apparent correlation with the coil library in side chain rotamers.[73]

Table 5 shows the relative free energies of α_L, α_R, PPII, and β conformers when the side chains adopt g+, g−, and t rotamers, from both our calculation and the statistical results. Not counting the conformers that are set as reference states with zero free energy or have too low a probability, the relative free energies calculated with our model are on average deviated from the statistical values by 1.98 kJ/mol. For the backbone conformers in the most favorable side chain rotamers, the average deviation is reduced to 1.12 kJ/mol.

Figure 7 shows the comparison of χ-dependent (ϕ, Ψ) Ramachandran plots between our model and the coil library for six representative amino acids. It is known that the shape and positions of the important minima on the (ϕ, Ψ) maps are different at the three χ conformers (g+, g−, and t).[73,78] For example, in the g+ conformers for all amino acids, ϕ in the left region spans the range of −150° to −60°, but the lower limit of ϕ extends to −180° in the g− conformers, expect for β-branched amino acids such as Val and Thr. Unlike the g+ and g− conformers, the upper limit of Ψ in the t conformers is lowered from 180° to 150°. In addition, in the g+ conformers, the α_L basins are well sampled, but they are normally less well sampled in the g− and t conformers. Our model is able to capture most of these features for various amino acids.

The χ-dependent Ramachandran plots of other amino acids are given in the SI (Figure S2). The results of our force field are in good agreement with the coil library results in each case.

Revision of Hydration Parameters of Charged Side Chains and Amide Groups. Globular proteins are soluble because there are polar side chains on the protein surface exposed to the hydrophilic environment. Although a protein core is normally composed of nonpolar groups, there are a significant number of polar or ionizable groups buried or partially buried inside a protein. Adolfsen et al. analyzed 124 PDB structures ranging from 100 amino acids to more than 600 amino acids.[79] About 37−61% of the surface area of ionizable groups is buried for 100-amino-acid-long proteins. Gunner et al. studied 490 proteins with a size of 36 to 1357 residues.[80] Among them, about 17% of the ionizable groups are fully buried in protein cores. Ionizable groups may not be charged if they are not exposed to water. For instance, Val66 of staphylococcal nuclease is buried in the core of the protein. When it is mutated to Lys, Asp, or Glu, the pK_a value of Lys is reduced to 5.7 and those of Asp and Glu are increased to 8.7 and 8.8, respectively.[81,82] A similar effect is also observed when Leu38 is mutated to Asp or Glu.[83] Another example is Asp79 in ribonuclease Sα, which has a pK_a value of 7.4.[84] The Asp79Phe mutant has a greater stability than the wide-type enzyme by 3.7 kcal/mol. This shows that ionizable amino acids may prefer a neutral state when they are placed in a hydrophobic environment. This can be physically understood since charged groups inside the protein induce a large desolvation penalty. During the folding process of a protein, ionizable residues would be neutralized before they move into the protein core. It may be particularly important to include this neutralization process in the force field in order to study the folding of large proteins that contain more buried ionizable groups.

In our previous work, we optimized the hydration parameters of charged groups by fitting the experimental hydration free energy of ions (ΔG_{hyd}).[41] In that parametrization scheme, we assumed that all ionizable side chains are always charged in both water and a low dielectric medium. In fact, ionizable groups may become neutral when they are in protein cores which normally have low dielectric constants. To account for this effect, in our model, we designed a thermodynamic cycle, as shown in Figure 8. Our goal is to attain an effective ΔG_{hyd} for an ionizable group so that regardless of the charge state of this group, its partition between water and cyclohexane agrees with experimental data. As such, hydration parameters of ionizable groups are reoptimized to fit the effective ΔG_{hyd}, which is calculated according to eq 7.

$$\Delta G_{hyd} = \Delta G_{chx} + \Delta G_{(hyd-chx)} + \Delta G_{pK_a} \qquad (7)$$

Effective solvation free energy in water (ΔG_{hyd}) is the sum of the experimental solvation free energy in cyclohexane (ΔG_{chx}),

Table 5. Relative Free Energies (kJ/mol) of α_L, α_R, PPII, and β Conformers of All Amino Acids except for Ala and Gly When Side Chains Adopt g+, g−, and t Conformations[a]

		coil library				PACE			
	side chain conformer	α_L	β	PPII	α_R	α_L	β	PPII	α_R
Lys	g+	3.0	1.0	0.0	0.0	5.0	3.2	0.2	0.0
	g−	11.0	0.0	2.0	2.0	8.4	0.0	4.3	2.9
	t	12.0	3.0	0.0	2.0	11.4	0.7	0.0	1.1
Gln	g+	4.0	1.0	0.0	0.0	5.7	3.2	0.1	0.0
	g−	10.0	0.0	2.0	3.0	n/a[b]	0.0	3.9	2.3
	t	10.0	3.0	0.0	1.0	11.4	0.2	0.0	1.2
Glu	g+	5.0	1.0	0.0	0.5	5.2	3.2	0.0	0.0
	g−	n/a	0.0	1.0	1.0	n/a	0.0	3.1	1.9
	t	11.0	4.0	0.0	1.0	10.4	0.4	0.0	1.6
Arg	g+	4.0	0.0	0.0	1.0	4.4	2.0	0.0	−0.3
	g−	12.0	0.0	2.0	4.0	10.9	0.0	4.3	2.2
	t	12.0	2.5	0.0	1.0	11.2	0.2	0.0	0.9
Met	g+	4.0	2.0	0.0	2.0	4.5	3.2	0.0	−0.1
	t	12.0	2.0	0.0	2.0	10.4	0.8	0.0	1.0
Leu	*g*+	7.0	2.0	0.0	2.0	6.7	4.2	0.0	1.6
	t	10.0	2.0	0.0	2.0	13.1	1.7	0.0	2.9
Val	g+	10.0	0.0	0.0	3.0	13.7	0.2	0.0	7.6
	g−	14.0	0.0	2.0	3.0	14.2	0.0	−0.3	3.7
	t	n/a	0.0	1.0	3.0	n/a	0.0	3.4	8.3
Ile	g+	11.0	0.0	0.0	4.0	14.7	0.0	0.1	6.7
	g−	13.0	0.0	3.0	3.0	n/a	0.0	0.2	3.7
	t	n/a	1.0	0.0	2.0	n/a	−3.5	0.0	4.6
Phe	g+	4.0	0.0	0.0	3.0	7.3	0.0	−1.3	3.2
	g−	n/a	0.0	4.0	5.0	n/a	0.0	5.6	8.4
	t	7.0	3.0	0.0	2.0	13.1	−0.5	0.0	6.9
Trp	g+	6.0	1.0	0.0	2.0	8.1	1.3	0.0	4.6
	g−	n/a	0.0	4.0	5.0	n/a	0.0	5.7	7.4
Tyr	g+	5.0	0.0	0.0	2.0	9.8	0.0	−0.7	3.8
	g−	n/a	0.0	4.0	5.0	13.8	0.0	5.9	12.1
	t	8.0	3.0	0.0	3.0	n/a	−0.7	0.0	6.1
His	g+	2.0	0.0	0.0	2.0	4.1	1.9	0.0	1.8
	g−	8.0	0.0	2.0	3.0	10.6	0.0	4.6	4.0
	t	4.0	1.0	0.0	1.0	9.4	−1.7	0.0	3.6
Cys	g+	4.0	2.0	0.0	3.0	6.6	2.5	0.0	0.4
	g-	n/a	0.0	3.0	1.0	n/a	0.0	2.3	2.2
Ser	g+	4.0	2.0	0.0	1.0	5.2	4.2	0.0	−0.2
	g−	13.0	1.0	1.0	0.0	10.4	4.0	2.0	0.0
	t	8.0	0.0	0.0	6.0	8.1	0.5	0.0	2.1
Thr	g+	8.0	2.0	0.0	2.0	11.9	0.0	0.0	6.8
	g−	11.0	0.0	2.0	0.0	10.1	0.0	−0.7	1.3
	t	17.0	0.0	3.0	7.5	n/a	0.0	3.9	6.2
Asn	g+	0.0	3.0	0.0	1.0	1.7	4.6	0.0	−0.8
	g−	11.0	0.0	3.0	0.0	9.1	3.9	2.5	0.0
	t	0.0	2.0	0.0	3.0	5.9	0.1	0.0	1.6
Asp	g+	3.0	4.5	0.0	1.0	2.5	7.4	0.0	−0.7
	g−	15.0	3.0	5.0	0.0	n/a	4.7	1.4	0.0
	t	2.0	2.0	0.0	3.0	5.1	0.0	0.0	0.2
Pro		n/a	n/a	0.0	4.0	n/a	n/a	0.0	3.3

[a] Both the results from the coil library and our force field are listed. [b] The probability of the conformer is too low for free energy to be calcualted. Standard deviations were estimated from block averages with a block size of 10 ns. The standard deviations for all conformers are less than 1.0 kJ/mol except for α_L conformers (<2.5 kJ/mol).

the experimental transfer free energy from cyclohexane to water ($\Delta G_{(hyd-chx)}$), and the free energy change of ionization in water (ΔG_{pK_a}). Experimental values of ΔG_{chx} and $\Delta G_{hyd-chx}$ are taken from the works of Radzicka and Wolfenden[85] and Wolfenden et al.[86] The ΔG_{pK_a} value can be obtained from eq 8.

$$\Delta G_{pK_a} = -RT \ln K_{eq}$$
$$\ln K_{eq} = 2.30(\mathrm{pH} - \mathrm{p}K_a) \quad (8)$$

R is the Boltzmann constant; T is 300 K. The pH is assumed to be 7. Experimental pK_a values of ionizable groups are taken from the works of Thurlkill et al.[87] and Nozaki and Tanford.[88] Solvation free energy values and revised parameters of ionizable groups are shown in Tables 6 and 7, respectively.

It should be noted that the hydration parameters of our ionizable groups are optimized in such a way that these groups can partition in water and a low dielectric medium as if they have distinct charge states in different environments. However, the parameters of nonbond interactions between these groups and other protein sites actually cannot vary in different environments.

As explicit hydrogen sites in side chain amide groups are included in the current model, the hydration scheme of this group is different from the previous model, in which the

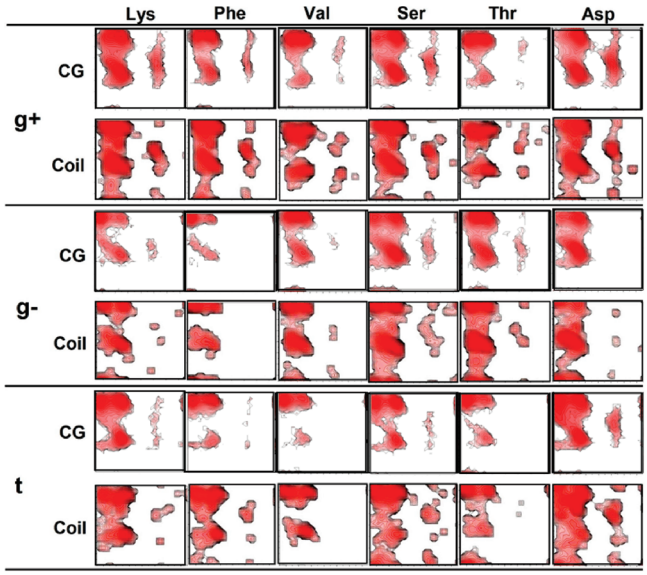

Figure 7. Backbone (ϕ, Ψ) distributions with side chains adopting different conformers (g+, g−, and t) for Lys, Phe, Val, Ser, Thr, and Asp from our UA model in CG solvent (CG) and the statistical results (coil). The gap between the contour lines indicates the free energy difference of 1 kJ/mol.

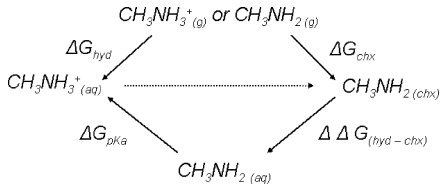

Figure 8. Revised hydration parametrization scheme for ionizable groups: Lys, Arg, Asp, and Glu. To account for the effect of the charged groups becoming neutral when they are in the protein core, the solvation free energy in water (ΔG_{hyd}) of ionizable groups is considered the sum of the solvation free energy in cyclohexane (ΔG_{chx}), the transfer free energy from cyclohexane to water ($\Delta G_{(hyd-chx)}$), and the free energy change of charging in water (ΔG_{pK_a}). Hydration parameters are optimized to fit ΔG_{hyd}.

Table 6. Calculated Solvation Free Energy of Ionizable Groups in Water (ΔG_{hyd}) from the Experimental Solvation Free Energy in Cyclohexane (ΔG_{chx}),[85,86] the Transfer Free Energy from Cyclohexane to Water ($\Delta G_{(hyd-chx)}$),[85,86] and the Free Energy Change When Charged (ΔG_{pKa})[a]

	ΔG_{hyd}	ΔG_{chx}	$\Delta G_{hyd-chx}$	ΔG_{pK_a}	pK_a
Lys	−37.3	−16.4	−1.5	−19.4	10.40[a]
Arg	−73.3	−20.6	−24.2	−28.5	12.0[b]
Asp	−47.0	−9.2	−18.8	−19.0	3.67[a]
Glu	−42.8	−13.9	−13.2	−15.7	4.25[a]

[a] All free energy values are in kJ/mol. pK_a values are taken from the work of Thurlkill et al.[87] and Nozaki and Tanford.[88]

hydrogen sites are implicitly represented. Thus, we also reoptimized the hydration parameters for interactions between CG water and the amide group. For the NH moiety of aromatic rings in His and Trp, its hydration parameters were transferred from the side chain amide group. As explicit amide hydrogen sites are also considered for backbone amide, the hydration parameters for backbone amide need a reoptimization. We find that most of the parameters for side chain amide can be transferred to backbone amide except for the hydration parameters of hydrogen sites, which were separately optimized.

The hydration parameters for charged side chains and amide groups are obtained by fitting the experimental hydration free energies of organic compounds like we did before.[41] All of the reoptimized parameters and calculated hydration free energies are listed in Table 7.

Potential of Mean Force of Polar Side Chains. It is important to accurately describe polar/charged interactions in order to study protein folding and protein−protein interactions. PMF calculation is a fundamental measure of these interactions in a water solvent. To accurately reproduce salt bridge and hydrogen bond interactions, interaction parameters of polar groups are optimized to fit PMFs obtained from all-atom simulations. As all nonbonded interactions in our model involve only two types of parameters, ε_{ij} representing interaction strength and δ_{ij} representing interaction distance, ε_{ij} and δ_{ij} parameters are optimized simultaneously in the PMF calculations. Optimized parameters for polar−polar interactions are shown in Table 8.

Figure 9 shows the PMFs of salt bridge or hydrogen bond interactions of eight polar side-chain pairs in water. As multidimensional PMF is still computationally challenging, 1D PMFs are obtained with the position restraint of polar groups on a straight line. The constrained orientations of polar groups are also shown in Figure 9. In our model, hydrogen(s) is implicitly incorporated into ammonium nitrogen, guanidinium nitrogen, and hydroxyl oxygen. Figure 9a shows the PMF of a typical salt bridge normally found in proteins between ammonium and carboxylate groups. This pair of groups was constrained by the collinear approach in that the $-CH_2NH_3^+$ moiety of ammonium and the $-CH_2-C<$ moiety of carboxylate were on the same line. It has an interaction energy of about −10.4 kJ/mol in our model, which is comparable to the all-atom simulation (−10.0 kJ/mol). The generalized Born (GB) model developed by Brooks et al. in 1999 underestimated the interaction energy by 3.3 kJ/mol, whereas EEF1 overestimated it by 6.3 kJ/mol.[89] Figure 9b shows the PMF of the guanidine−carboxylate interaction. This salt bridge pair was fixed on the same plane, and the $-NH-C<$ moiety of guanidine and the $-CH_2-C<$ moiety of carboxylate were constrained on a line. Our model is able to reproduce the interaction energy, which is about −18.4 kJ/mol in our force field and −18.8 kJ/mol in the all-atom model. The GB model is 2.5 kJ/mol less attractive, and EEF1 is about 17 kJ/mol too attractive.[89] We assume that HB acceptor groups such as amide carbonyl groups and acceptor nitrogen in aromatic rings share the same parameters with carboxylate groups when they interact with donor groups such as ammonium, amide, and hydroxyl. In addition, the parameters for like-charge groups are optimized in a similar manner, the PMFs of which are shown in Figure 9c−d.

Solvent-separated interactions can be observed in explicit solvent models only because this requires a single water molecule to be placed between two polar groups. Although this feature is apparent in our force field, the location of the solvent-separated interaction is shifted further away by about

Table 7. Optimized Parameters for Interactions between CG Water and Charged Side Chains or Amide Groups and the Calculated and Experimental ΔG_{hyr} of the Compounds That Are Used to Fit the Parameters[a]

particle type[b]	ε_{ij} (kJ/mol)	δ_{ij} (nm)	compounds	ΔG_{hyr} (kJ/mol)	ΔG_{hyr}(expt) (kJ/mol)
$-\underline{NH_3}^+$	12.00	0.340	butylammonium	−37.4	−37.3
$-\underline{NH}-C-\underline{(NH_2)_2}$	4.90	0.340	N-propylguanidinium	−73.1	−73.3
$-NH-\underline{C}-(NH_2)_2$	4.90	0.340			
$-C\underline{OO}^-$	6.00	0.340	acetate ion	−45.8	−47.0
$-\underline{C}OO^-$	2.00	0.415			
$-CO-\underline{N}H_2$	0.86	0.400	acetamide	−40.3	−40.6
$-C\underline{O}-NH_2$	9.00	0.340			
$-\underline{C}O-NH_2$	0.8	0.415			
$-CO-N\underline{H_2}$	3.40	0.280			
$-CO-N\underline{H}-$	7.00	0.280	N-methylacetamide	−42.4	−42.1

[a] The CG simulations are at 300 K. The calculated ΔG_{hyr} is an average over 6−8 simulations. The standard deviations are less than 1.5 kJ/mol. [b] $-NH_3^+$ for ammonium; $-COO^-$ for carboxylate; $-CO-NH_2$ for side chain amide; $-CO-NH-$ for backbone amide; $-NH-C-(NH_2)_2$ for guanidine.

Table 8. Optimized Parameters for Polar−Polar Interactions of Charged/Polar Groups

interacting particles[a]	ε_{ij}	δ_{ij} (nm)
$-\underline{NH_3}^+\cdots-C\underline{OO}^-$	10.00	0.260
$-\underline{NH_3}^+\cdots-C\underline{O}-NH_2$	10.00	0.260
$-\underline{NH_3}^+\cdots=\underline{N}-$	10.00	0.260
$-\underline{NH}-C-\underline{(NH_2)_2}\cdots-C\underline{OO}^-$	10.00	0.260
$-NH-\underline{C}-(NH_2)_2\cdots-\underline{C}OO^-$ [b]	1.00	0.360
$-\underline{NH}-C-\underline{(NH_2)_2}\cdots-C\underline{O}-NH_2$	3.00	0.300
$-CO-N\underline{H_2}\cdots-C\underline{OO}^-$	23.00	0.160
$-CO-\underline{N}H_2\cdots-C\underline{OO}^-$ [c]	15.00	0.235
$-CO-N\underline{H_2}\cdots-\underline{C}OO^-$ [c]	15.00	0.235
$-CO-N\underline{H_2}\cdots-C\underline{O}-NH_2$	23.00	0.160
$-CO-\underline{N}H_2\cdots-C\underline{O}-NH_2$[c]	15.00	0.235
$-CO-N\underline{H_2}\cdots-\underline{C}O-NH_2$[c]	15.00	0.235
$-CO-N\underline{H_2}\cdots=\underline{N}^-$	23.00	0.160
$-CO-\underline{N}H_2\cdots=\underline{N}^-$ [c]	15.00	0.235
$-\underline{OH}\cdots-C\underline{OO}^-$	7.20	0.260
$-\underline{OH}\cdots-C\underline{O}-NH_2$	7.20	0.260
$-\underline{OH}\cdots-\underline{OH}$	6.20	0.260
$-\underline{OH}\cdots=\underline{N}^-$	7.20	0.260
$-CO-N\underline{H}-\cdots-C\underline{OO}^-$	27.00	0.16
$-CO-N\underline{H}-\cdots-C\underline{O}-NH_2$	27.00	0.16
$-\underline{NH_3}^+\cdots-\underline{NH_3}^+$ [b]	4.00	0.33
$-C\underline{OO}^-\cdots-C\underline{OO}^-$ [c]	2.00	0.28
$-\underline{NH}C\underline{(NH_2)_2}\cdots-\underline{NH}C\underline{(NH_2)_2}$[b]	4.00	0.33

[a] $-NH_3^+$ for ammonium; $-COO^-$ for carboxylate; $-CO-NH_2$ for side chain amide; $-CO-NH-$ for backbone amide; $-NH-C-(NH_2)_2$ for guanidine; $=N-$ for acceptor nitrogen in heterocycle; $-OH$ for hydroxyl group. [b] The following potential is used for this interaction: $E_{nonbond} = \sum_{i\neq j}(4\varepsilon_{ij}\delta_{ij}^6)/r^6$. [c] The following potential is used for this interaction: $E_{nonbond} = \sum_{i\neq j}(4\varepsilon_{ij}\delta_{ij}^{12})/r^{12}$.

2 Å (Figure 9). This is because each of our CG solvent particles represents four water molecules. In addition, the solvent-mediated interactions are examined only for pairs of small molecules. Further studies are needed to investigate these interactions involved with much larger molecules such as protein−protein interactions. Thus, we should be cautious in studying proteins that involve significant solvent-separated interactions. Like implicit solvent models, the height of the first PMF peak is underestimated in our force field. However, it is not clear whether there is any significant consequence in capturing such fine details of the interactions. While the folding kinetics might be altered, the lower height of the first PMF peak might in fact accelerate the conformation sampling without introducing much thermodynamic bias.

In our model, polar hydrogen is explicitly represented in amide (Asn, Gln), methylimidazole (His), and methylindole (Trp) groups. With explicit hydrogens, hydrogen bonds

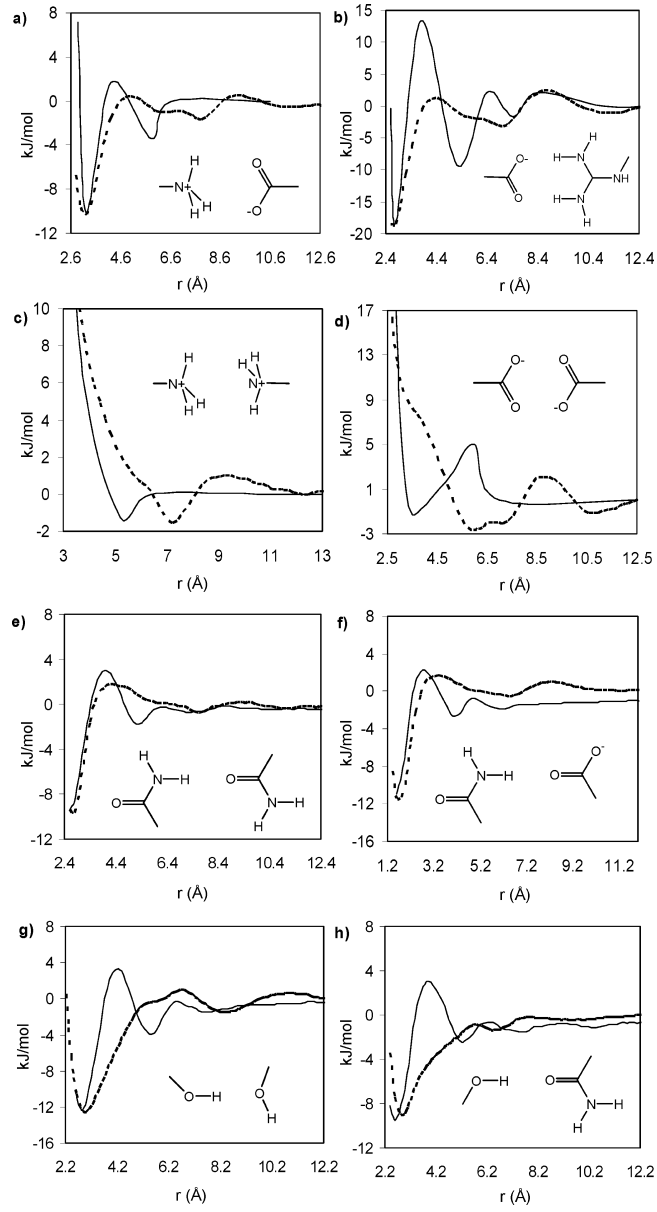

Figure 9. PMFs of six polar group pairs in the UA model (dashed lines) and all-atom model (solid lines). *r* is the distance between atoms of two groups. The all-atom PMFs in a−d come from ref 89. In g and h, the value of 10.0 kJ/mol is used for ε_{attr}.

involving these polar groups could have better orientation and directionality. Here, we use the same parameters for the NH moiety in these different donor groups. As an example, amide/amide and amide/carboxylate PMFs are shown in Figure 9e−f. Two polar groups are constrained on the same plane in the PMF calculation. The hydrogen bond interaction energy of amide/amide is −9.4 and −9.5 kJ/mol in our model and the all-atom model, respectively. The amide/carboxylate pair also shows comparable results to the all-atom simulation. The contact energy in the all-atom model is −9.2 kJ/mol and is −9.4 kJ/mol in our model.[63] As the backbone NH moiety has a larger hydration parameter than the NH moiety in side chain amide (Table 7), we also optimized its interaction parameters with the acceptors like carbonyl and carboxylate groups.

As a hydroxyl group is treated as one site currently, the inclusion of both its donor and acceptor ability may cause trouble when it simultaneously interacts with multiple donor and acceptor groups. We here are only focused on its HB donor property, except for the case of hydroxyl/hydroxyl interactions. A statistical analysis of the PDB structures showed that a hydroxyl group is 3 times more likely to be a donor than an acceptor.[90] PMFs of hydroxyl/hydroxyl and hydroxyl/amide hydrogen bond interaction are shown in Figure 9g and h, respectively. For all hydrogen bond cases, polar group pairs are constrained so that the hydrogen bond donor, the hydrogen atom, and the hydrogen bond acceptor are on the same line. The hydroxyl/hydroxyl pair has about −12.6 and −12.3 kJ/mol of contact energy in our force field and the all-atom model, respectively. The hydrogen bond interaction energy of the hydroxyl/amide pair is −9.0 kJ/mol in our model and about −9.5 kJ/mol with the all-atom model. As the hydrogen site is implicit in our force field, the potential derived in the above corresponds to a situation in which the OH bond is always in the best orientation to form a HB when it is close to an acceptor group. However, because the OH bond can rotate freely, placing it in the HB direction can cause a loss of entropy. We considered this effect by adding a correction of $RT \ln(1/3)$ onto the attraction parameters between hydroxyl and acceptor groups due to the alignment of the OH bond in the HB direction from three rotamer states. Similarly, we added $RT \ln(1/3 \times 2/3)$ for the hydroxyl/hydroxyl interaction. Therefore, the original attraction parameter (ε) of 10.0 kJ/mol for hydroxyl/acceptor and hydroxyl/hydroxyl now becomes 7.2 and 6.2 kJ/mol, respectively.

Optimization of Backbone/Backbone HB. Backbone/backbone HB interactions are critical for defining the geometrical feature of proteins, especially for secondary structures. They are the most abundant interaction type in proteins. As expected, any small variation in the parameters of backbone/backbone interactions can significantly influence the simulation results of a force field. Unlike the parametrization of side chain amide groups, we tried to optimize the backbone/backbone HB parameters through peptide simulations. Such a strategy has been used before. For instance, Brooks and co-workers refined backbone potentials and atomistic input radii for continuum electrostatics in the development of their CHARMM-CAMP force field.[63] Irback and Mohanty used a series of folding simulations of peptides and mini-proteins to optimize nonbonded interactions of their all-atom force field with an implicit solvent model.[91] These studies raise the importance of balancing between an α helix and β sheet. We here chose two peptides, the AK17 [Ac-(AAKAA)$_3$GY-NMe][92] peptide that is ∼30−35% helical and the GB1m2 (Ac-GEWTYNPATGKFTVTE-NMe)[93] peptide that is a mutant of the N-terminal β-hairpin of the protein G B1 domain. The extremely fast kinetics of these peptides enable us to perform numerous folding simulations to optimize the backbone/backbone HB parameters. Each folding simulation was conducted in 16 REMD replicas with a temperature ranging from 300 to 430 K. Each simulation lasts for 100−200 ns. The average helical and sheet contents were calculated over the last 50 ns of the simulation.

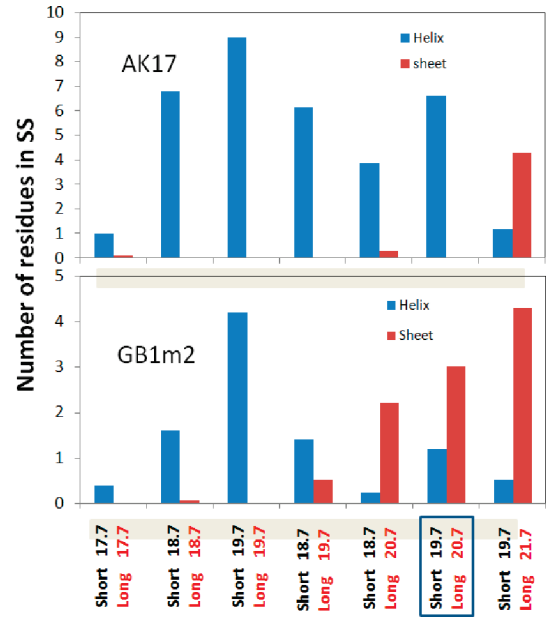

Figure 10. Number of helical (blue) and sheet (brown) residues of the AK17 and GB1m2 peptides with different ε_{attr} parameters (kJ/mol). "Short" indicates the values of ε_{attr} for the short-range O_i/N_{i+3} and O_i/N_{i+4} HBs. "Long" indicates the values of ε_{attr} for the other HBs.

Equation 5 and Figure 2a show all of the potential terms of the backbone/backbone HBs. Only the attraction parameter ε_{attr} that controls HB strength was further optimized. It should be noted that no attempts have been made to reproduce the native structures of the peptides. Only the total contents of the secondary structures of the peptides are fitted in the optimization. Figure 10 shows the secondary structure contents of AK17 and the GB1m2 with different ε_{attr} values. With ε_{attr} increasing from 17.7 to 19.7 kJ/mol, the helical content of AK17 increases from ∼7% to ∼61%. However, instead of a β sheet, a significant amount of α helix is also developed in the GB1m2, indicating that a single ε_{attr} value is not enough. We therefore adopt two ε_{atrr} parameters. One is for short-range hydrogen bonds between O_i and N_{i+3} and between O_i and N_{i+4} that are crucial to turns and helices. The other is for the remaining long-range HB bonds. After a nontrivial trial-and-error optimization, we find that ε_{attr} is about 19.2 kJ/mol for the short-range HBs and ε_{attr} is about 20.7 kJ/mol for the long-range HBs.

Table 9. Comparison between the Averaged C_α RMSD (nm) from the Respective Experimental Structures for the Nine Proteins Simulated with Our Force Field and Previous All-Atom Models[a]

PDB ID		no. of AAs	UA		All-atom	
			simulation length (ns)	C_α RMSD (nm)	simulation length (ns)	C_α RMSD (nm) [b]
3gb1	NMR	56	100	0.258 (0.029)	50	0.106 (0.020)
1ctf	X-ray	68	100	0.242 (0.045)	100	0.16[c]
1d3z	NMR	76	100	0.259 (0.026)	22	0.141 (0.020)
1bta	NMR	89	100	0.415 (0.101)	142.9	0.134 (0.016)
1fks	NMR	107	100	0.248 (0.017)	143.5	0.358 (0.074)
1fw7	NMR	108	100	0.381 (0.052)	148.0	0.171 (0.015)

[a] The values in parentheses are standard deviations. Average RMSD values were calculated from the beginnings of simulations. [b] These are from ref 50. [c] This value is from ref 95 and is an average over the last 90 ns.

Aqueous Simulations of Proteins. Like all-atom force fields, our force field is built upon the parametrization of small molecules such as side chain analogues or dipeptides that contain constituent functionality. There remains a question of whether it is effective and accurate to combine all of the optimized parameters for the study of real proteins. A stringent test should be the folding of peptides and mini-proteins into their native structures by first principle means. Such a test is presented in our accompanying paper. Here, we show a preliminary test on protein simulations. Aqueous simulations of proteins have become a consensual way of examining the protein portions of force fields.[94] Using this approach and starting with NMR/X-ray structures, multi-nanosecond MD simulations of proteins in water have been conducted. During the simulations, the maintenance of experimental structures as well as other thermodynamic properties is deemed an indication of the feasibility of force fields for protein simulations.

Verification of Our Force Field. To examine our force field, we selected seven proteins of small to medium size (56−108 AA) that have a cross-section of various secondary structures. Most of them have recently been used to examine the performance of a modified version of the CHARMM force fields.[95] Another (PDB code: 1ctf) was used to validate the modification of the GROMOS96 force fields.[50] The RMSD values of Cα carbon atoms from native structures were employed to monitor whether the native structures are kept during the simulations. Table 9 lists the average C_α RMSD values in 100 ns simulations of the six proteins with our force field. The full trajectories of the RMSD can be found in Figure S3 in the SI. All of the calculated RMSDs are around or below 0.3 nm except for 1bta. The average RMSD for all of the proteins is about 0.295 nm. Thus, the native structures can be considered maintained. Our results are generally larger than but comparable to the RMSD of the same proteins derived from all-atom simulations. In one case (1fks), our RMSD is even better than that of all-atom models (0.248 nm vs 0.358 nm). In terms of the ability to reproduce experimental native structures, our force field is roughly comparable to the all-atom models.

Finally, we examined whether our force field can be extended into more complex systems such as large proteins. We chose to study malate synthase G (PDB code: 2jqx), which contains 723 amino acids.[96] Starting with the NMR-derived structure, a 30 ns MD simulation was carried out. Throughout the simulations, the RMSD is kept around 0.4 nm (Figure 11a). The average RMSD is 0.43 nm. Considering the large size of the protein, this RMSD is at least small enough as an indication of the maintenance of the native topology (Figure 11b). It should be noted that the simulation was conducted with a 2.66G Hz dual-core CPU at a speed of 10 ns/day. Combining the solvents, our force field system is equivalent to an all-atom system of more than 100 000 atoms. Simulating tens or hundreds of nanoseconds of such systems is still a formidable task for all-atom force fields.

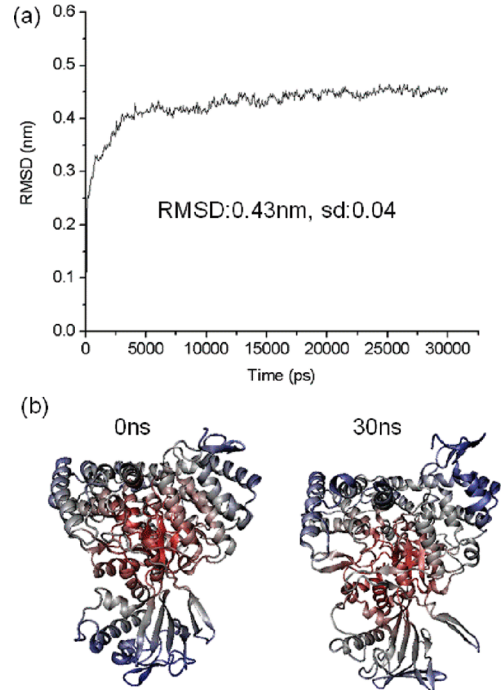

Figure 11. (a) Change of the root-mean-square deviation (nm) of C_α carbons during the simulation. (b) The PDB structure of malate synthase G and its simulated structure at the end of the simulation.

We do notice that several proteins including 1bta and 1fw7 partially unfold during the simulations (Table 9 and Figure S3, SI). Visual inspection reveals that the unfolding is mainly caused by a considerable loss of native polar interactions such as HBs. We suspect that our force field may underestimate polar interactions in a low dielectric medium such as a large protein. This is because all of the parameters for polar interactions are optimized to reproduce all-atom results in a solvent-exposed environment. In a low dielectric medium,

the polar interactions are supposed to be strengthened due to the removal of the screening effect of water. Nevertheless, the screening effect is missing in the current CG model that is basically LJ fluid. Thus, our force field may be more suitable for proteins with large proportions of exposed parts.

Conclusion

In this paper, we describe our recent effort in the improvement of our protein force field model, which includes two aspects. First, the backbone (ϕ, Ψ) potential is reoptimized to reproduce shapes, positions, and depths of β, PPII, α_R, and α_L basins on the statistical (ϕ, Ψ) surfaces for Ala and Gly. Local side chain−backbone interactions are optimized to reproduce correct preferences of side chain conformers (g+, g−, and t) of various amino acids as well as the dependence of backbone conformations (ϕ, Ψ) on side chain conformations (χ). Second, we also parametrize the interactions between polar groups by fitting the PMFs of polar group pairs that are derived from all-atom simulations. Together with previously derived parameters, our force field is able to reproduce the native structures of series of proteins of small to medium size (<150 AA). Moreover, the stability of the native structures can be maintained for even larger proteins (>700 AA). Simulating tens of nanoseconds of such large proteins is easy (several days of calculation with a single computer) for our force field, and simulations of microseconds are within reach. These results imply the applicability of our force field to the structural study of real proteins.

Acknowledgment. We are grateful to the Research Grants Council of Hong Kong (CA06/07.SC05, 663509) for financial support of this research.

Supporting Information Available: Tables S1−S5, the parameters of our force field; Figure S1, the equilibrium values of bond lengths and bond angles of our force field; Figure S2, the χ-dependent Ramachandran plots of Gln, Glu, Arg, Met, Leu, Ile, Trp, Tyr, His, Cys, and Asn; and Figure S3, the plots of RMSDs in all six protein simulations. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chen, J. H.; Brooks, C. L.; Khandogin, *J. Curr. Opin. Struct. Biol.* **2008**, *18*, 140.

(2) van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. *Curr. Opin. Struct. Biol.* **2008**, *18*, 149.

(3) MacKerell, A. D.; Nilsson, L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 194.

(4) Schaeffer, R. D.; Fersht, A.; Daggett, V. *Curr. Opin. Struct. Biol.* **2008**, *18*, 4.

(5) Sanbonmatsu, K. Y.; Tung, C. S. *J. Struct. Biol.* **2007**, *157*, 470.

(6) Villa, E.; Balaeff, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6783.

(7) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694.

(8) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945.

(9) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534.

(10) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859.

(11) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623.

(12) Sippl., M. J. *J. Comput.-Aided Mol. Des* **1993**, *7*, 473.

(13) Matysiak, S.; Clementi, C. *J. Mol. Biol.* **2006**, *363*, 297.

(14) Khurana, E.; DeVane, R.; Kohlmeyer, A.; Klein, M. L. *Nano Lett.* **2008**, *8*, 3626.

(15) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750.

(16) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812.

(17) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819.

(18) Bond, P. J.; Holyyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593.

(19) Treptow, W.; Marrink, S. J.; Tarek, M. *J. Phys. Chem. B* **2008**, *112*, 3277.

(20) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123.

(21) Aksimentiev, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4337.

(22) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806.

(23) Miyazawa, S.; Jernigan, R. L. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 357.

(24) Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646.

(25) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697.

(26) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849.

(27) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874.

(28) Yap, E. H.; Fawzi, N. L.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 626.

(29) Khatun, J.; Khare, S. D.; Dokholyan, N. V. *J. Mol. Biol.* **2004**, *336*, 1223.

(30) Dokholyan, N. V. *Curr. Opin. Struct. Biol.* **2006**, *16*, 79.

(31) Chebaro, Y.; Dong, X.; Laghaei, R.; Derreumaux, P.; Mousseau, N. *J. Phys. Chem. B* **2009**, *113*, 267.

(32) Maupetit, J.; Tuffery, P.; Derreumaux, P. *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 394.

(33) Fujitsuka, Y.; Chikenji, G.; Takada, S. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 381.

(34) Takada, S.; Luthey-Schulten, Z.; Wolynes, P. G. *J. Chem. Phys.* **1999**, *110*, 11616.

(35) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469.

(36) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116.

(37) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289.

(38) Ding, F.; Tsao, D.; Nie, H. F.; Dokholyan, N. V. *Structure* **2008**, *16*, 1010.

(39) Ding, F.; Buldyrev, S. V.; Dokholyan, N. V. *Biophys. J.* **2005**, *88*, 147.

(40) Han, W.; Wu, Y.-D. *J. Chem. Theory Comput.* **2007**, *3*, 2146.

(41) Han, W.; Wan, C.-K.; Wu, Y.-D. *J. Chem. Theory Comput.* **2008**, *4*, 1891.

(42) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511.

(43) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10.

(44) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144.

(45) Klein, M. L.; Shinoda, W. *Science* **2008**, *321*, 798.

(46) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.

(47) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(48) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781.

(49) MacKerell, A. D.; Fieg, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400.

(50) Cao, Z.; Lin, Z.; Wang, J.; Liu, H. Y. *J. Comput. Chem.* **2009**, *30*, 645.

(51) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712.

(52) Hu, H.; Elstner, M.; Hermans, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 451.

(53) Kaminski, G.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(54) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908.

(55) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.

(56) Head-Gordon, T.; Head-Gordon, M.; Frisch, M. J.; Brooks, C. L., III; Pople, J. A. *J. Am. Chem. Soc.* **1991**, *113*, 5989.

(57) Brooks, C. L., III; Case, D. A. *Chem. Rev.* **1993**, *93*, 2487.

(58) Cornell, W. D.; Cieplk, P.; Bayly, C.; Gould, I. R.; Merz, K. M. J.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *11*, 5179.

(59) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Feig, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-MacCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.

(60) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657.

(61) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4* (5), 819.

(62) Makowski, M.; Sobolewski, E.; Czaplewski, C.; Ołdziej, S.; Liwo, A.; Scheraga, H. A. *J. Phys. Chem. B* **2008**, *112*, 11385.

(63) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728.

(64) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.

(65) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463.

(66) Berendsen, H. J. C.; Pstma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(67) Feenstra, K. A.; Hess, B.; Berendsen, H. J. C. *J. Comput. Chem.* **1999**, *20*, 786.

(68) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.

(69) Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001**, *335*, 435.

(70) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474.

(71) Bogusz, S.; Cheatham, T. E.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070.

(72) 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.

(73) Jiang, F.; Han, W.; Wu, Y.-D. *J. Phys. Chem. B* **2010**, *114*, 5840.

(74) Jha, A. K.; Colubri, A.; Zaman, M. H.; Koide, S.; Sosnick, T. R.; Freed, K. F. *Biochemistry* **2005**, *44*, 9691.

(75) Lovell, S. C.; Davis, I. W.; Arendall, W., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437.

(76) Garcia, A. E.; Sanbonmatsu, K Y. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 2782.

(77) Ho, B. K.; Thomas, A.; Brasseur, R. *Protein Sci.* **2003**, *12*, 2508.

(78) Chakrabarti, P.; Pal, D. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1.

(79) Kajander, T.; Kahn, P. C.; Passila, S. H.; Cohen, D. C.; Lehtio, L.; Adolfsen, W.; Warwicker, J.; Schell, U.; Goldman, A. *Structure* **2000**, *8*, 1203.

(80) Kim, J.; Mao, J.; Gunner, M. R. *J. Mol. Biol.* **2005**, *348*, 1283.

(81) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; Garcýa-Moreno, E. B. *Biophys. J.* **2002**, *82*, 3289.

(82) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; Garcýa-Moreno, E. B. *Biophys. J.* **2007**, *92*, 2041.

(83) Harms, M. J.; Castañeda, C. A.; Schlessman, J. L.; Sue, G. R.; Isom, D. G.; Cannon, B. R.; García-Moreno, B. *J. Mol. Biol.* **2009**, *389*, 34.

(84) Trevino, S. R.; Gokulan, K.; Newsom, S.; Thurlkill, R. L.; Shaw, K. L.; Mitkevich, V. A.; Makarov, A. A.; Sacchettini, J. C.; Scholtz, J. M.; Pace, C. N. *J. Mol. Biol.* **2005**, *354*, 967.

(85) Radzicka, A.; Wolfenden, R. *Biochemistry* **1988**, *27*, 1664.

(86) Wolfenden, R.; Anderson, L.; Cullis, P.; Southgate, C. *Biochemistry* **1981**, *20*, 849.

(87) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2006**, *15*, 1214.

(88) Nozaki, Y.; Tanford, C. *Methods Enzymol.* **1967**, *11,* 715.

(89) Masunov, A.; Lazridis, I. *J. Am. Chem. Soc.* **2003**, *125*, 1722.

(90) Eswar, N.; Ramaakrishnan, C. *Protein Eng.* **2000**, *13*, 227.

(91) Irback, A.; Mohanty, S. *Biophys. J.* **2005**, *88*, 1560.

(92) Luo, P.; Baldwin, R. L. *Biochemistry* **1997**, *36*, 8413.

(93) Feinmeyer, R. M.; Hudson, F. M.; Anderson, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238.

(94) Price, D. J.; Brooks, C. L., III. *J. Comput. Chem.* **2002**, *23*, 1045.

(95) Feig, M. *J. Chem. Theory Comput.* **2008**, *4*, 1555.

(96) Grishaev, A.; Tugarinov, V.; Kay, L. E.; Trewhella, J.; Bax, A. *J. Biomol. NMR* **2008**, *40*, 95.

CT1003127