A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations

YONG DUAN,[1] CHUN WU,[1] SHIBASISH CHOWDHURY,[1] MATHEW C. LEE,[1] GUOMING XIONG,[1] WEI ZHANG,[1] RONG YANG,[1] PIOTR CIEPLAK,[2,3] RAY LUO,[2] TAISUNG LEE,[2,3] JAMES CALDWELL,[2] JUNMEI WANG,[2] PETER KOLLMAN[2,†]

[1]*Department of Chemistry and Biochemistry, University of Delaware, Newark, Delaware 19716*

[2]*Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, California 94143*

[3]*Accelrys Inc., 9685 Scranton Rd, San Diego, California 92121*

Received 7 April 2003; Accepted 1 July 2003

Abstract: Molecular mechanics models have been applied extensively to study the dynamics of proteins and nucleic acids. Here we report the development of a third-generation point-charge all-atom force field for proteins. Following the earlier approach of Cornell et al., the charge set was obtained by fitting to the electrostatic potentials of dipeptides calculated using B3LYP/cc-pVTZ//HF/6-31G** quantum mechanical methods. The main-chain torsion parameters were obtained by fitting to the energy profiles of Ace-Ala-Nme and Ace-Gly-Nme di-peptides calculated using MP2/cc-pVTZ//HF/6-31G** quantum mechanical methods. All other parameters were taken from the existing AMBER data base. The major departure from previous force fields is that all quantum mechanical calculations were done in the condensed phase with continuum solvent models and an effective dielectric constant of $\varepsilon = 4$. We anticipate that this force field parameter set will address certain critical short comings of previous force fields in condensed-phase simulations of proteins. Initial tests on peptides demonstrated a high-degree of similarity between the calculated and the statistically measured Ramanchandran maps for both Ace-Gly-Nme and Ace-Ala-Nme di-peptides. Some highlights of our results include (1) well-preserved balance between the extended and helical region distributions, and (2) favorable type-II poly-proline helical region in agreement with recent experiments. Backward compatibility between the new and Cornell et al. charge sets, as judged by overall agreement between dipole moments, allows a smooth transition to the new force field in the area of ligand-binding calculations. Test simulations on a large set of proteins are also discussed.

 J Comput Chem 24: 1999–2012, 2003

Key words: point-charge force field; quantum mechanical calculations; molecular mechanics simulations

Introduction

Advancements in the field of molecular mechanics simulations have concentrated in the areas of simulation methodologies in the past. Some highlights of these include accurate treatment of electrostatics,[1] efficient conformational sampling methods,[2,3] methods for massively parallel simulations,[4,5] and continuum solvent models.[6,7] These developments have enabled long-time simulations close to the folding time of small proteins,[4] and several successful folding simulations of miniproteins have been reported recently.[8–11] Thus, the conformational sampling ability of all-atom models has reached an important threshold at which simulations of many biologically relevant processes are increasingly routine; however, the development of force fields has lagged behind. With the growing interest in ever more realistic simulations, the need for a reasonably accurate and efficient force field that can represent macromolecules in the condensed phase is becoming more critical.

In molecular mechanics-based molecular dynamics simulations, the molecular systems are represented by molecular mechan-

Correspondence to: Y. Duan; e-mail: yduan@udel.edu

[†]Deceased

Contract/grant sponsor: NIH Research Source; contract/grant number: CRR-15588

Contract/grant sponsor: National Institute of General Medical Sciences; contract/grant numbers: GM-64458 and GM-67168 (to Y.D.), and GM-29072 (to P.A.K.)

Contract/grant sponsor: The State of Delaware and University of Delaware Research Fund (to Y.D.)

ics models in which the parameters are developed based on fundamental physical principles. The accuracy of a simulation is largely determined by two factors: conformation sampling, and model accuracy. In the past, the conformational sampling ability of molecular dynamics simulation was viewed as the primary bottleneck.[12,13] With the advancements in simulation methodologies[14,15] and the increase in computer speed, this limitation is gradually diminishing. Increasingly, accuracy of the underlying models becomes the dominant factor in the outcome of a simulation. For example, with current simulation methodologies and fast computers one can readily reach the folding time scales of small peptides using continuum solvent models,[16] and in some cases, using all-atom solvent models[4]; however, whether the simulations can accurately reflect physical reality now depend on the force field parameters used to represent the physical interactions.

An important characteristic of the current molecular mechanics models is that their parameters are obtained through high-level quantum mechanical calculations on short peptide fragments. Such an approach has several advantages. It assures generality and allows further refinement upon the availability of more accurate quantum mechanical methods. Because the parameters are integral components of a molecular mechanics model that describes the interacting molecular forces, they are often referred to as the "force field" parameters. Among the early all-atom force field developers, Weiner et al. successfully adopted this approach[17] and developed one of the first-generation molecular mechanics force fields based primarily on quantum mechanical data. Due to the inadequate computing power of the time, much of the applications of this force field were limited to *in vaccuo* simulations.[18] A decade later, after considerable accumulation of simulation data, Cornell et al.[19] developed one of the second-generation force fields based on improved quantum mechanical calculations. Since then, the Cornell et al. force field has enjoyed a wide range of applications in simulating both nucleic acids and proteins. One remarkable early success of Cornell et al. force field was that it demonstrated the ability to move incorrect conformations to the correct ones when combined with accurate calculation of electrostatic forces.[20,21]

Despite their successes, the existing "second-generation" force fields are still based on gas-phase quantum mechanical calculations, simulations of small molecules, and typically nanosecond time-scale simulations of proteins. Recent developments in force fields include the fully polarizable force fields from Kollman group[22] and from Friesner group.[23] These force fields can potentially provide accurate representations in both the gas phase and the condensed phase, and is suited for simulations in a variety of solvent environments (e.g., membrane proteins). On the other hand, these force fields all tend to sacrifice computational efficiency for accuracy. This factor alone may discourage their widespread use. An even serious drawback for the inclusion of polarizability is the reduced integration stability. Although the computational overhead can be minimized to 30–50%, our tests show that a 0.5-fs integration time step is required to obtain a stable trajectory (Duan, unpublished results). In comparison, a fourfold larger 2.0-fs time step is typically used in conventional all-atom simulations. A sixfold reduction in overall efficiency is expected when the computational overhead is taken into account. Furthermore, because of the multibody interactions in the polarizable force field, its application to Monte Carlo simulations is far from being straightforward.

In contrast to the polarizable formulation, the traditional all-atom representation of force fields strikes a balance between performance and accuracy. However, until recently, all-atom force fields have been developed based on gas-phase quantum mechanical calculations of small peptides. One major deficiency in this approach is that such a force field lacks proper representation of the polarization effect in condensed phase. For example, the dipole moment of alanine-dipeptide in extended conformation is 1.51 Debye in the gas phase and 1.74 Debye in organic solvent ($\varepsilon = 4$), differing by more than 15%. Consequently, force fields developed based on gas-phase quantum mechanical calculations underestimate the electrostatic interactions when applied to the study of proteins. In reality, they are more suited for small peptides in the gas phase rather than proteins in the condensed phase. In the rest of this article, we will describe our new atom-centered point charge all-atom force field that was design to address this deficiency.

Methods

In keeping with our previous minimalist approach, this model is an effective two-body additive model. The potential function describing interactions among particles comprises electrostatic, van der Waals, bond, bond angle, and dihedral terms [eq. (1)]

$$E_{\text{total}} = \sum_{\text{bonds}} K_b(b - b_{eq})^2 + \sum_{\text{angle}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right] \quad (1)$$

where, K_b and K_θ are the force constants for the bond and bond angles, respectively; b and θ are bond length and bond angle; b_{eq} and θ_{eq} are the equilibrium bond length and bond angle; ϕ is the dihedral angle and V_n is the corresponding force constant; the phase angle γ takes values of either 0° or 180°. The nonbonded part of the potential is represented by van der Waals (A_{ij}) and London dispersion terms (B_{ij}) and interactions between partial atomic charges (q_i and q_j). ε is the dielectric constant that takes into account of the effect of the medium that is not explicitly represented and usually equals to 1.0 in a typical solvated environment where solvent is represented explicitly. The nonbonded terms are calculated for all atom pairs that are either separated by more than three bonds or are not bonded. Interactions between atoms separated by three bonds account for the one to four interactions in which the electrostatic and van der Waals parts are reduced by 20–50%, depending on the specific implementation of the force field. In this version, the one to four electrostatic interactions are divided by a factor of 1.20 and the Lennard–Jones terms are divided by 2.0. Both scaling factors are identical to the Cornell et al. force field for consistency.

At the present stage of development, we have chosen to rederive the charges and refit the main-chain torsion parameters only because we believe that these two factors contribute most signif-

icantly to the accuracy of the model. The main-chain torsion parameters are especially crucial in maintaining a reasonable balance between the important secondary conformations because errors could be cumulative in protein models. Other parameters, including bond, bond angle, side-chain torsion, and van der Waals parameters are retained from the existing AMBER force field parameter set.[19,24]

Charge Derivation

In the Cornell et al. force field, the effective charges were obtained by fitting the gas-phase electrostatic potential of small peptides calculated by HF/6-31G* using RESP.[25] Because the charges derived using such an approach systematically overestimate dipole moments, much like the charges in the point-charge water models, they implicitly include the solvent polarization effect to some extent. The polarization effect is important in the condensed phase where the local electrostatic environment is significantly different than that in the gas phase due to the presence of neighboring atoms. In the present approach, the electrostatic potential is calculated in the condensed phase with continuum solvent model and the effective point-charges are obtained by RESP fitting. Furthermore, the advancement in computational methodology allows us to calculate the electrostatic potentials more accurately using the DFT method with large basis set.

The new charges were obtained by fitting to the quantum mechanically derived electrostatic potentials using B3LYP/cc-pVTZ//HF/6-31G** methods. In these calculations, each amino acid was represented by a dipeptide fragment consisting of the amino acid residue and the terminal groups (Ace- and -Nme). The electrostatic potentials of each peptide were calculated for two conformations with main-chain dihedral angles constrained to (Φ, Ψ) = (-60, -40) and (Φ, Ψ) = (-120, 140), respectively, representing the α-helical and the extended conformations. The initial conformations were generated using AMBER simulation package with a simulated-annealing protocol. These dipeptide conformers were then subjected to energy minimization using the AMBER Cornell et al. force field. Further geometry optimizations were done at the RHF/6-31G** level of QM theory. All QM calculations were done using the Gaussian 98 simulation package.[26] Single-point calculations were done using the density functional theory (DFT) method and the B3LYP exchange and correlation functionals[27–29] with the ccpVTZ[30] basis set. The IEFPCM continuum solvent model[31,32] was applied to mimic an organic solvent environment ($\varepsilon = 4$). The electrostatic potentials of the solutes (peptides) were saved and were used in the charge fitting.

Effective charges were obtained by fitting the electrostatic potential of peptides using RESP method.[25] A two-stage fitting procedure was used. In the first stage of fitting, the two conformers of each dipeptide were combined. In the second stage, the chemically equivalent atoms were set to have the same charges, while the charges of the terminal blocking groups and those of heavy atoms were fixed. Because the charge-matching process may introduce small errors, we purposely limited the errors to the matched atoms by fixing the charges of the blocking groups and heavy atoms. Finally, the charges of the blocking groups were fit by combining electrostatic potentials of all amino acids.

In the Cornell et al. charge set,[19] the main-chain charges were determined by a combined RESP fit of different dipeptides such that the charged amino acids have the same main-chain charges while the others share another set of main-chain charges. In this work, we chose to let each amino acid have its own main-chain charges. Intuitively, this should allow sequence-dependent features to be incorporated into the main-chain charge set.

Main-Chain Torsion Parameters

The main-chain torsion parameters, C–N–C_α–C, N–C_α–C–N, C–N–C_α–C_β, and N–C–C_α–C_β, were obtained by fitting to a 2D (Φ–Ψ) 144-point energy profile of alanine-dipeptide calculated using the MP2/cc-pVTZ QM method and the IEFPCM continuum solvent model[33] with a dielectric constant of $\varepsilon = 4$ after restrained geometry optimization with RHF/6-31G**. These 144 points are on a 2D grid of the Φ–Ψ torsion angles with a grid size of 30° (or 12 points) in each direction. This allowed us to fit the third Fourier term of the main-chain torsion parameters, including all four torsion angles listed above (or a total of 12 parameters). Boltzmann's weighting factors, $w = \exp(-0.2E)$, were used to ensure that the high degree of difference is mainly localized in the energetically unfavorable regions (E is the QM energy in kcal/mol). Overall a weighted [by $\exp(-0.2E)$] RMS difference of 1.7 kcal/mol was obtained.

The torsion parameters, C–N–C_α–H_α and N–C–C_α–H_α of Glycine, were obtained by fitting to a 36-point energy profile that is equally distributed on the 2D Φ–Ψ grid. Each of these torsion parameters were calculated up to the second Fourier term. This fitting procedure differed significantly from our earlier approach where the main-chain torsion parameters were obtained by fitting to a few key conformers of blocked Alanine dipeptide. Despite the excellent fitting results in our earlier attempts, both versions[19,34] of the force field were found to be biased toward either α-helical conformation (in the earlier version) or β-extended conformation (in the later version).

It is important to note that the purpose of the fitting is to ensure that the energetic surface of Ace-Ala-Nme is adequately represented. However, comparisons between the quantum and molecular mechanical data of this particular peptide have been a common practice in the field to judge the accuracy of the force field. Thus, the accuracy of Ace- and -Nme charges used in the fitting of torsion parameters becomes an important issue. This, of course, does not reflect their significance in protein modeling where, in fact, they are rarely used. Here we choose to use the charges of Ace- and -Nme of the Ala dipeptide to reduce the likelihood of over compensation. We also choose not to match the charges of those chemically equivalent atoms of the Ace- and -Nme groups to maximize the accuracy of these charges, and hence, the underlying electrostatic potential they represent. The final reported charges of the same groups were obtained by a combined fit of all dipeptides. The fitted torsion parameters are given in Table 1.

Molecular Dynamics Simulations

Molecular dynamics simulations were conducted on a number of peptides including Ace-Gly-Nme, Ace-Ala-Nme, and Ace-Ala_4-Nme (Ala_4) in explicit solvent represented by the TIP3P model.[35]

Table 1. Main-Chain Torsion Parameters.

	$\nu1$	$\gamma1$	$\nu2$	$\gamma2$	$\nu3$	$\gamma3$
N–C_α–C–N	0.6839	180	1.4537	180	0.4615	180
C–N–C_α–C	1.0159	0	0.3451	180	0.2259	0
N–C–C_α–C_β	0.7784	180	0.0657	180	0.0560	0
C–N–C_α–C_β	0.3537	180	0.8836	180	0.2270	180
[a]C–N–C_α–H_α	0.4575	0	1.2558	180		
[a]N–C–C_α–H_α	0.5607	180	0.0110	0		

The symbols follow those in Eq. (1). Only those torsion parameters that are different from earlier version of AMBER force field[24] are given. Up to three Fourier terms are used for main-chain torsions, whereas only two Fourier terms are used for C–N–C_α–H and N–C–C_α–H of Glycine.
[a]C–N–C_α–H and N–C–C_α–H are for Gly only. $\nu_{1,2,3}$ are in kcal/mol and $\gamma_{1,2,3}$ are in degrees.

These simulations were performed to examine the accuracy of the parameters and the balance between the important conformations of peptides in aqueous solution. The Ala_4 peptide is one of the smallest peptides that can potentially form two main-chain hydrogen bonds. In these simulations, the peptides were initially in the extended conformations. A 100-ps simulation was conducted at 800 K using Generalized Born continuum solvent model[7] to randomize the initial conformation. Solvent molecules were then added around the peptides in truncated octahedral periodic boxes. The minimum distances from the peptide atoms to the surfaces of the boxes were set to 10 Å, with a total of approximately 4510 atoms (or 1500 water molecules). The simulations were started by short (500 steps) energy minimizations and the initial velocities were assigned randomly with a Gaussian distribution at $T = 100$ K. The temperatures were raised to 300 K over 10 ps, and were maintained at 300 K using a Berendsen thermostat.[36] The simulations continued for more than 8.0 ns. The subsequent analysis was based on the later part of the simulation after excluding the initial 100-ps equilibration phase. Particle Mesh Ewald[1] was used to treat the long-range electrostatic interactions and the Lennard–Jones interactions were truncated at 8.0 Å. A time step of 2.0 fs was used in the simulations. Pressure was maintained at 1.0 pa using Berendsen algorithm, and the periodic boundary condition was imposed by both minimum image and the Particle Mesh Ewald.

Ramanchandran Contour Maps of Peptides

The potential of mean force (PMF) contour maps of the main-chain Φ–Ψ distribution were constructed from high-resolution X-ray crystallography structures. The nonhomologous chains were selected by PISCES.[37] The selection criteria included lower than 40% sequence homology, 2.5 Å or better resolution, and smaller than 0.25 *R*-factors. A total of 2150 chains and 432,576 residues were selected with these criteria. Histograms were made by statistical sampling of the main-chain Φ and Ψ torsion angles of the residues on a 12×12 grid (30° intervals in each direction). The histograms were converted to PMF maps at 300 K by the formula $G = -0.597 \ln(n)$ (kcal/mol), where n is the occurrence at the grid point. The maps were shifted such that the lowest free energy is zero. Contours were made at 1.0 kcal/mol intervals after second-order spline interpolation.

Results

Extensive tests were conducted to assess the accuracy of the new force field. In particular, its ability to represent both extended and helical regions in a balanced manner were closely scrutinized based on comparisons with both quantum mechanical data and the PMF obtained from high resolution X-ray protein structures. In keeping with traditions of the force field development community, the outline of our results presentation will be as follows: We will first present the comparisons with quantum mechanical data on both Ace-Ala-Nme and Ace-$(Ala)_4$-Nme peptides. We then present comparisons with the Cornell et al. charge set and other charges calculated using a variety of quantum mechanical theories to assess the quality of the new charges and dipole moments followed by simulation results on three peptides their comparisons to PMF. Finally, we will conclude with our initial test results on other small peptides and on a large ensemble of decoy set.

Comparison with Quantum Mechanical Data

Almost all contemporary force field parameters have been developed based on comparison with relatively high-level quantum mechanical data. This is a generally accepted practice, particularly because quantitative experimental data on short peptides is still difficult to obtain. For example, Friesner and coworkers have recently refined the OPLS-AA force field based on the LMP2/cc-pVTZ(-f) data.[38] Here we have also followed the conventional practice by comparing our results against the quantum mechanical data.

Figure 1 shows the QM and MM energies (in kcal/mol) of Ace-Ala-Nme. The QM energies were calculated using MP2/cc-pVTZ//HF/6-31G** method in $\varepsilon = 4.0$ medium. The MM energies were calculated in the "gas phase" after constrained energy minimization with the main-chain torsion angles fixed at the designated values. Because our charges were derived from QM data in $\varepsilon = 4.0$ medium, our "gas phase" charges effectively mimic such environment and is consistent with the QM energies under comparison. Overall, the (unweighted) root-mean-square difference between the QM and MM energies is 1.9 kcal/mol, and the average absolute difference is 1.4 kcal/mol. The energy RMSD in the regions of the Ramanchandran plot relevant to the typical protein conformations is 0.565 kcal/mol, and the average absolute difference is 0.48 kcal/mol. These regions are defined by the 5.0 kcal/mol contour line of the MP2 energy map (shown in Fig. 1). This is comparable to the level of accuracy obtainable from the quantum mechanical method (MP2/cc-pVTZ//HF/6-31G**)[39] and is considered acceptable. Notable differences include the slight shift of the minimum around the α-helical region and the somewhat expanded and slightly more favorable contour lines in the same area, suggesting that the new force field has a tendency to over-represent the helical region, which is contradictory to simulation results in water (discussed later).

Another difference is seen on the $\Phi > 0$ side of the Φ–Ψ map around the α_L region where the new force field appears to over-

Table 2. Comparison between the QM and MM Energies of the tetra-Ala Peptide.

Conformer		QM	MM	Φ_1	Ψ_1	Φ_2	Ψ_2	Φ_3	Ψ_3
	1	0.00	0.49	−146	157	−145	160	−145	156
	2	0.13	0.71	−159	164	−155	158	−86	79
$\Phi < 0$	4	1.42	1.29	−156	162	−89	84	−157	153
	5	1.17	0.61	−157	170	−78	−18	−155	161
	α_R	5.69	5.31	−52	−53	−52	−53	−52	−53
	RMSD		0.46						
	3	−2.71	−3.03	−82	92	76	−53	−81	85
	6	−0.51	−0.12	−89	67	64	24	−166	151
$\Phi > 0$	7	3.06	1.51	56	−159	−93	64	−163	−50
	8	1.45	2.88	73	−71	−58	135	62	26
	9	4.21	1.90	76	−58	76	−56	76	−55
	10	4.28	5.50	62	30	65	21	74	−52

Energies are in kcal/mol. Main-chain torsion angles are also given in the table for reference. Corresponding main-chain torsion angles after energy minimization using the new force field are within 10° except the α_R conformation in which the torsion angles were restrained to the QM geometry. The QM data has been provided by Friesner.[39] Energies are in kcal/mol, and angles are in degrees.

represent the region by about 2.0 kcal/mol. However, our simulation on Ace-Ala-Nme peptide, the same peptide from which QM data was obtained, clearly showed that the α_L region was not sampled at all during the simulation of the peptide in solution (discussed in detail later). In our opinion, merely fitting to QM data cannot guarantee an accurate force field. Indeed, it probably makes sense to fit to gas-phase QM data if the molecular mechanical force field is polarizable. On the other hand, for point-charge models, fitting to the gas-phase QM data would make the model more gas phase like. This is quite undesirable because of the differences in the electrostatic potentials in the gas phase and in the condensed phase. A more serious problem is that if the charge set is condensed phase like, such as the Cornell et al. charge set, and this charge set, fitting the main-chain torsion parameters to the gas-phase QM data, would result in over compensation, which in turn, produces a set of parameters with a potential bias towards a particular conformation (either α- of β-) when used in condense-phase protein simulations. Because of the uncertain and varying dielectric environment, QM data, although quantitative, should serve as merely a guide, and the accuracy of the force field should be better judged by simulations of small peptides, even though the latter is often qualitative. Motivated by this observation, we devoted significant efforts to extended simulations on small peptides. The results of these simulations are discussed later.

Beachy et al.[39] compared the QM energies of 10 Ace-$(Ala)_3$-Nme tetra-peptide conformers, calculated with LMP2/cc-pVTZ-(-f)//HF/6-31G** QM methods, to the MM energies calculated by various force fields. Our comparisons to the same set of energies are summarized in Table 2. Because tetra-peptide is the smallest peptide that can form one main-chain hydrogen bond in α-helical conformation, we have also included the α-helical conformer (courtesy of Friesner) in our comparison. This test served as the critical set for evaluating the energetic balance between α- and β-conformations. Our result gave an estimate for the molecular mechanical energy difference between α- and β-conformations, which is 4.8 kcal/mol (Table 2), and is in reasonable agreement with the quantum-mechanically calculated energy difference between α- and β-conformations (5.7 kcal/mol).

Among the 11 conformers studied by Beachy et al.,[39] conformers 7 and 9 exhibited the largest differences relative to conformer 1 (1.9 and 2.6 kcal/mol, respectively). Conformer 9 is a left-handed helix, which occurs only infrequently in proteins. These results suggest that the MM energy surface overrepresented the left-handed helical conformations by more than 0.8 kcal/mol per residue, consistent with our earlier assessment based on the 2D Φ–Ψ energy maps. The readers should be aware that these results are in contradiction to the simulation results of both Ace-Ala-Nme and Ace-$(Ala)_4$-Nme peptides in solvent (discussed later). In either case, this disagreement with the QM energy is not a major concern because of the low frequency of occurrences in protein structures. When conformers containing $\Phi > 0$ (conformers 3, 6, 7, 8, 9, and 10) were excluded, the root-mean-square difference was 0.46 kcal/mol. Geometrically, all Φ and Ψ torsion angles were within 10° from the QM-optimized geometry after unrestrained energy minimization with the MM force field, except the α-helical conformer, which was restrained to the QM geometry (data not shown).

It should be noted that the QM energies of the tetra-Ala were obtained in the gas phase. Therefore, caution must be taken when they are compared to MM energies of point-charge models. Because gas-phase energies can described only the behavior of small peptides in the gas phase (despite their perceived accuracy), it would be misleading to use them to compare against force field parameters that are designed to mimic proteins in solution. It is interesting that the difference between these two environments has been neglected even in some of the recent studies. Our own studies clearly indicated significant changes on the energetic surfaces of small peptides at different dielectric environments. In particular, the extended conformation, which corresponds to the β-sheet secondary structure, is considerably more favorable in the condensed phase due to solvent polarization (data not shown). Such a change should have profound effect on the balance between helical

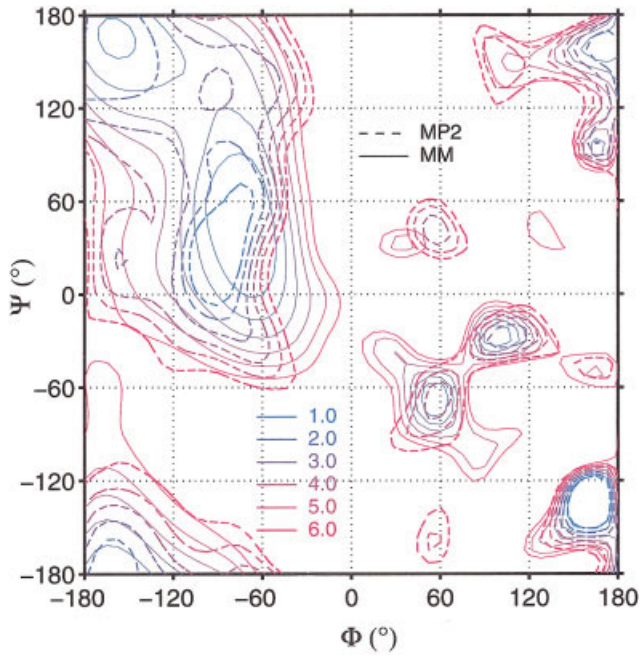

Figure 1. Comparison between the QM and MM energy maps. Energies were calculated on a 12 × 12 grid. A second-order spline interpolation was applied to obtain smooth contours. Contour lines are drawn at 1.0-kcal intervals, starting from 1.0 kcal. Solid contour lines represent the MM energy map and dashed lines are for QM. QM energies are calculated using MP2/cc-pVTZ//HF/6-31G** in $\varepsilon = 4.0$ medium.

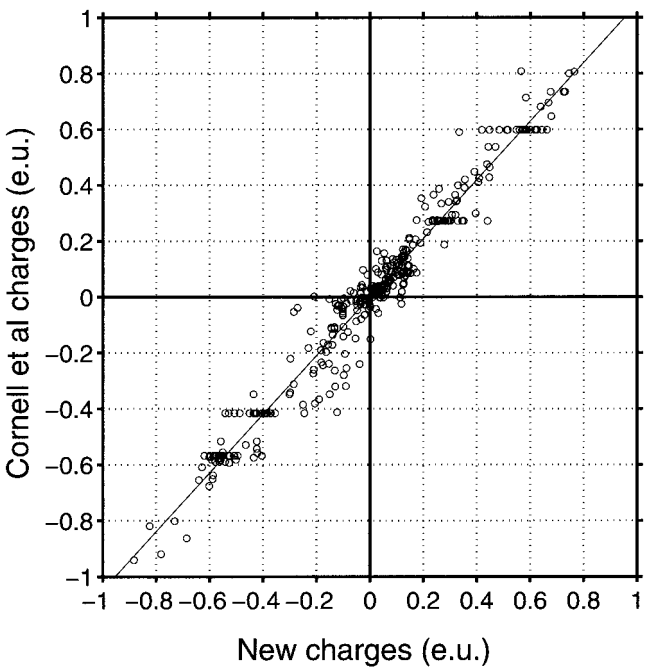

Figure 2. Comparison between the new charge set and the Cornell et al. charge set. The trendline has a slope of 1.048.

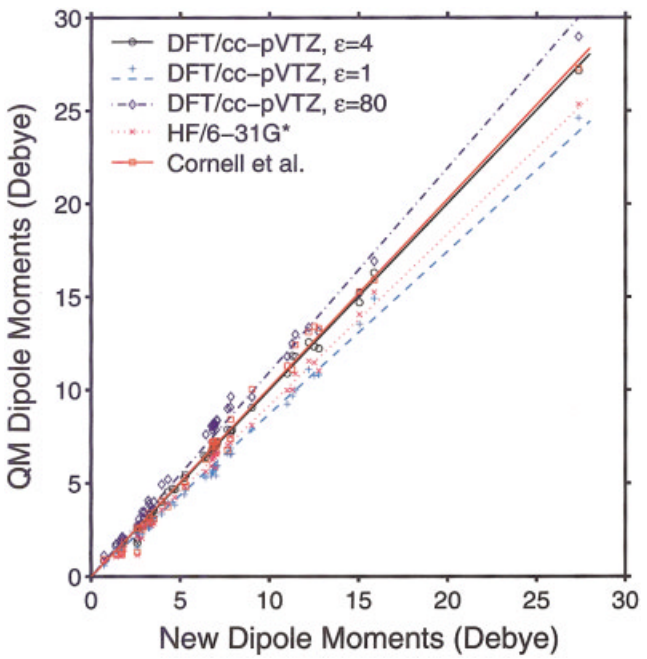

Figure 3. Calculated dipoles are compared for the standard amino acids (constructed as Ace-X-Nme) with identical sets of structures. The DFT calculations were done using B3LYP with cc-pVTZ basis set in the specified media. Comparisons to the dipoles moments calculated using HF/6-31G* and Cornell et al. charge set are also presented. The slopes of the trendlines are given in Table 5.

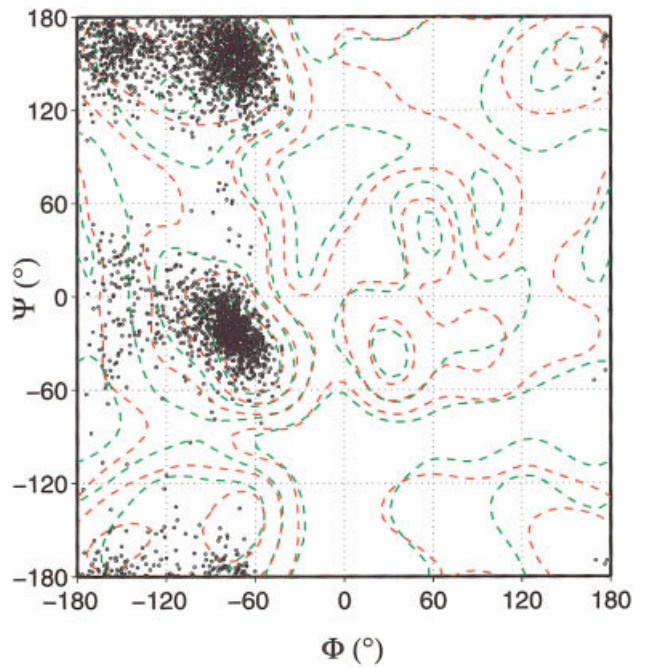

Figure 4. Ramanchandran plot of Ace-Ala-Nme di-peptide from simulation in TIP3P water and is compared to the PMF contours obtained from statistical analysis of high-resolution X-ray structures. Please see refs. 44 and 45 for comparison with other force fields. The red contours are the PMF of Ala and the green contours are for all residues except Gly and Pro.

Table 3. Atomic Partial Charges (in e.u.) of Standard Amino Acids.

	Gly	Ala	Ser	Cys	Val	Thr	Pro	Ile	Leu	Met	Asp	Asn	Glu	Gln	His	Lys	Arg	Trp	Phe	Tyr
N	−0.374	−0.405	−0.541	−0.396	−0.450	−0.245	−0.088	−0.451	−0.355	−0.395	−0.558	−0.430	−0.423	−0.387	−0.528	−0.436	−0.301	−0.428	−0.371	−0.488
H	0.254	0.294	0.345	0.295	0.440	0.255		0.329	0.262	0.281	0.320	0.255	0.307	0.301	0.282	0.251	0.234	0.242	0.234	0.264
C	0.581	0.570	0.483	0.643	0.447	0.560	0.334	0.569	0.573	0.600	0.443	0.617	0.470	0.419	0.662	0.725	0.730	0.584	0.548	0.622
O	−0.509	−0.555	−0.581	−0.585	−0.405	−0.552	−0.435	−0.620	−0.558	−0.566	−0.501	−0.524	−0.593	−0.565	−0.529	−0.563	−0.578	−0.495	−0.507	−0.527
C_α	−0.129	−0.028	0.118	−0.074	−0.052	−0.271	−0.035	−0.102	−0.101	−0.088	0.007	0.045	0.032	0.037	0.031	−0.039	−0.131	−0.020	−0.030	0.010
[a]H_α	0.089	0.121	0.142	0.141	−0.026	0.164	0.060	0.174	0.137	0.123	0.082	0.060	0.065	0.152	0.085	0.129	0.053	0.107	0.102	0.096
C_β		−0.230	0.147	−0.221	0.395	0.238	−0.003	0.062	−0.144	0.019	−0.048	−0.094	0.075	−0.032	−0.152	−0.108	0.037	−0.098	−0.099	−0.052
[b]H_β		0.077	0.040	0.147	−0.116	0.045	0.019	0.062	0.053	0.049	−0.015	0.043	−0.004	0.031	0.055	0.045	0.028	0.065	0.061	0.019
[c]C_γ, O_γ, S_γ			−0.640	−0.285	−0.090	−0.602	0.013	0.022	0.192	−0.208	0.745	0.584	−0.034	−0.020	0.278	0.033	0.012	−0.100	0.021	0.113
[d]H_γ			0.446	0.189	−0.009	0.405	0.020	0.012	0.001	0.124			−0.004	0.031		0.010	0.003			
$C_{\gamma 2}$						−0.176		−0.130												
$H_{\gamma 2(1,2,3)}$						0.060		0.030												
[e]C_δ, O_δ, N_δ							−0.012	−0.101	−0.123	−0.212	−0.730	−0.527	0.765	0.668	−0.423	−0.048	0.126	−0.174	−0.083	−0.183
[f]H_δ							0.044	0.024	0.022							0.071	0.068	0.171	0.098	0.133
[g]$C_{\delta 2}$, $N_{\delta 2}$												−0.782			−0.298			0.090		
[h]$H_{\delta 2}$												0.355			0.160					
[i]C_ε, O_ε, N_ε										−0.285			−0.824	−0.628	0.026	−0.070	0.465	−0.298	−0.157	−0.182
[j]H_ε										0.128					0.127	0.120	0.326	0.322	0.124	0.137
[k]$C_{\varepsilon 2}$, $N_{\varepsilon 2}$														−0.883	−0.098			0.142		
[l]$H_{\varepsilon 2}$														0.408	0.267					
$C_{\varepsilon 3}$																		−0.154		
$H_{\varepsilon 3}$																		0.123		
[m]C_ζ																−0.250	0.566	−0.211	−0.100	0.206
[n]H_ζ																0.295		0.126	0.115	
[o]$C_{\zeta 3}$, O_H, $N_{H(1,2)}$																	−0.686	−0.164		−0.421
[p]$H_{\zeta 3}$, H_H																	0.391	0.119		0.330
C_{H2}																		−0.133		
H_{H2}																		0.119		

[a]$H_{\alpha(2,3)}$ for Gly.
[b]$H_{\beta(1,2,3)}$ for Ala and H_β for Thr, Ile, and Val, $H_{\beta(2,3)}$ for all others.
[c]C_γ for Glu, Asp, Lys, Pro, Met, Asn, and Gln; $C_{\gamma(1,2)}$ for Val; O_γ for Ser; $O_{\gamma 1}$ for Thr; S_γ for Cys.
[d]$H_{\gamma 1}$ for Thr, $H_{\gamma(2,3)}$ for Gln, Arg, $H_{\gamma 1(2,3)}$ for Ile, $H_{\gamma(1,2)(1,2,3)}$ for Val.
[e]$C_{\delta 1}$ for Ile; Trp; $C_{\delta(1,2)}$ for Leu, Phe, Tyr; S_δ for Met; $O_{\delta 1}$ for Asn; $O_{\delta(1,2)}$ for Asp; C_δ for Pro, Glu, Gln, Lys, Arg; $N_{\delta 1}$ for His.
[f]$H_{\delta 1(1,2,3)}$ for Ile, $H_{\delta(2,3)}$ for Arg, Lys, Pro; $H_{\delta 1}$ for Trp; $H_{\delta(1,2)}$ for Phe, Tyr; $H_{\delta(1,2)(1,2,3)}$ for Leu.
[g]$C_{\delta 2}$ for His, Trp; $N_{\delta 2}$ for Asn.
[h]$H_{\delta 2(1,2)}$ for Asn.
[i]C_ε for Met, Lys; $C_{\varepsilon 1}$ for His; $C_{\varepsilon(1,2)}$ for Tyr, Phe, $O_{\varepsilon 1}$ for Gln; $O_{\varepsilon(1,2)}$ for Glu; N_ε for Arg; $N_{\varepsilon 1}$ for Trp.
[j]$H_{\varepsilon(2,3)}$ for Lys; $H_{\varepsilon(1,2,3)}$ for Met; H_ε for Arg; $H_{\varepsilon 1}$ for His, Trp; $H_{\varepsilon(1,2)}$ for Phe, Tyr.
[k]$C_{\varepsilon 2}$ for Trp, Phe, Tyr; $N_{\varepsilon 2}$ for Gln, His.
[l]$H_{\varepsilon 2(1,2)}$ for Gln.
[m]$C_{\zeta 2}$ for Trp; N_ζ for Lys.
[n]$H_{\zeta 2}$ for Trp; $H_{\zeta(1,2,3)}$ for Lys.
[o]$C_{\zeta 3}$ for Trp; O_H for Tyr; $N_{H(1,2)}$ for Arg.
[p]$H_{\zeta 3}$ for Trp; H_H for Tyr; $H_{H(1,2)(1,2)}$ for Arg.

Table 4. Slopes of the Trend Lines That Best Fit the Data Shown in Figure 3.

	DFT/Gas	DFT/ε = 4	DFT/ε = 80	HF/6-31G*/Gas	Cornell et al.
Dipoles	0.872	1.002	1.097	0.919	1.013
Charges	0.950	—	1.038	1.102	1.049

The slopes of trend lines of the charges are also given in the table.

and β-sheet conformations. Thus, we consider the aforementioned 0.46 kcal/mol RMS energy difference among the selected tetra-Ala conformers acceptable.

Comparisons of the Charges and Dipole Moments

As we indicated earlier, an important feature of Cornell et al. charge set is that it is obtained by fitting to the electrostatic potential calculated using the HF/6-31G* quantum mechanical method. Because HF/6-31G* has a tendency to exaggerate the gas-phase dipole moment, the charges are slightly larger (in absolute value) than the typical gas-phase charges. In some sense, the charges are somewhat condensed phase-like. Because our new charge set was obtained by fitting to the condensed-phase electrostatic, it would be interesting to compare these two sets of charges. Shown in Figure 2 is the scatterplot of the charges. One can clearly see a good correlation between them. Indeed, the correlation coefficient is 0.98. This indicates that the two sets of charges are highly similar, which is expected. However, the fitted straight line has a slope of 1.049, indicating that the Cornell et al. charge set is systematically larger than the new charge set by approximately 5%. The difference is not uniform; the fitting error ranges from -0.2 to 0.2 e.u., which is substantial.

Further comparisons were made to three additional charge sets obtained by fitting directly to the quantum mechanical data, including HF/6-31G*, B3LYP/cc-pVTZ in the gas phase, and B3LYP/cc-pVTZ in water, represented by the COSMO model.[40–43] The results are summarized in Table 4. The gas-phase charges obtained by fitting to the HF/6-31G* data were about 16% larger than those obtained from B3LYP/cc-pVTZ gas-phase data, 10% larger than the new charge set, and 5% larger than those from the B3LYP/cc-pVTZ water data. This seems to indicate that the HF/6-31G* charges are more polar than the COSMO charges, which is contrary to the conclusion based on comparisons of dipole moments (discussed below). In comparison, the new charge set (Table 3) is about 5% larger than the gas-phase charges obtained from B3LYP gas-phase data and about 4% smaller than the COSMO charges. Therefore, the new charges are close to the middle point between the gas phase and the aqueous solution phase.

Dipole moments are important physical properties of molecules. In our opinion, they are more important than the partial charges when the formal charges of the molecules are known. In the cases of amino acids, because all formal charges are known and most are neutral, dipole moments become the most important terms to account for the long-range electrostatic interactions. For example, molecules with larger dipole moments tend to be more hydrophilic than those of smaller dipole moments.

Interestingly, the dipole moments derived from the new charge set are systematically (about 8%) larger than those calculated using HF/6-31G* method, which was the basis of the Cornell et al. charge set, even though the new charges are about 10% smaller than the HF/6-31G* charges (Table 4). This sounds somewhat puzzling. But further analysis indicates that, although the new charges are smaller, they distribute differently in reflection of the distribution of effective charges of the underlying electronic structures. Because the HF/6-31G* method exaggerates the electrostatic potentials, the resultant charges are somewhat larger than the gas-phase charges, as if the charges are simply scaled up in comparison to gas-phase charges. Such an effect can partially mimic the condensed-phase electrostatics. However, the electron distribution is expected to be polarized in the presence of solvent molecules. In particular, the polarization has a stronger effect on the surface atoms than on those buried atoms. Thus, surface atoms are more polarized than the buried atoms, and would have larger partial charges in comparison to the gas phase. The electrostatic potentials obtained from the condensed-phase quantum mechanical calculations are higher at the short range than those in the gas phase. This may not make much difference for small molecules, where almost all atoms are effectively exposed. For relatively large molecules, however, the difference can be substantial, as indicated from this comparison. Although uniform scaling charges can achieve the goal of increasing the dipoles, such an approach cannot mimic the true behavior of the molecules when they are surrounded by solvent particles.

Although the comparison with Cornell et al. charges suggests that the new charge set would make the peptides less hydrophilic, the dipole moments are comparable (shown in Fig. 3) when judged by the slope of the trendline which is 1.01 (show in Table 4). Interestingly, the Cornell et al. dipole moments are larger than the HF/6-31G*; therefore, the charge set is more polar than the underlying QM data. In fact, one may draw the conclusion that the dipoles based on Cornell et al. charge set are similar to those calculated in ε = 4 medium and the new charges, given that the slope of the fitted line is 1.01 (Table 4), even though Cornell et al. charge set is systematically larger than the new charge set by about 5%. Thus, the distribution of the new charges is different.

In comparison to other QM methods, the dipole moments calculated using the new charge set are comparable to those obtained using B3LYP/ccpVTZ in ε = 4 medium whose trendline has a slope of 1.002. They are about 10% smaller than those in water, about 13% larger than those in the gas phase, as represented by B3LYP/ccpVTZ. The dipole moments in water, as calculated using B3LYP/ccpVTZ with the COSMO continuum solvent

model, and $\varepsilon = 80$, are more than 20% greater than those in the gas phase, which is expected.

One of the obstacles of using HF/6-31G* to derive charges was that the HF/6-31G* level of theory lacks sufficient accuracy to be used in the torsion parameter refinement. A typical remedy was to use the other high-level QM theory (e.g., MP2) to obtain an accurate energy profile. A problem may arise due to the difference in the media in which the QM calculations were done. Because the charges are exaggerated due to HF/6-31G* electrostatic potentials, which mimics condensed-phase electrostatics, the torsion-fitting against the gas-phase QM data has the potential to over compensate. Although, this is acceptable for small peptides when the energies are compared to the gas-phase QM energies, the over compensation may become a source of error that has the potential to grow with increasing size of peptides. It could become a source of considerable error when the force field is applied to proteins or secondary structure fragments. Nevertheless, this was not a major concern in the past because the majority of applications were relatively short equilibrium simulations. Another obstacle is that the HF/6-31G* theory lacks the ability to predict the specific effect on the distribution of the charges when solvent and other atoms are present. Fortunately, the development of continuum solvent models has filled the gap.

Our present approach represents a step forward in force field development methodologies. The use of the continuum solvent models in the quantum mechanical calculations made it possible to represent the solvent polarization effects in the point-charge models in a systematic manner. This is an important feature, given the disparity between the gas-phase and condensed-phase electrostatic potentials. Thus, we would expect our parameter set to be able to mimic the condensed phase better. Consequently, our force field may be more suited for condensed-phase simulations by design. In particular, we speculate that our force field may be more accurate in modeling the side-chain and tertiary contacts, both of which are crucial for protein folding and other applications.

Finally, it is worthwhile to mention the increased level of QM calculation used in our development protocol. The partial charges were obtained by fitting the electrostatic potentials of peptides calculated using DFT quantum mechanical method and the B3LYP functionals[27–29] and the ccpVTZ basis set. This method is significantly more accurate than the typical HF/6-31G* level of theory in other force field development. In particular, the DFT method is significantly more accurate than the Hartree–Fock method typically used in other force fields (the latter lacks electron correlation). The increased accuracy can be directly translated into more accurate charges.

Simulation of Ace-Ala-Nme in TIP3P Water

In the past, one of the primary difficulties in developing protein force field was the lack of detailed quantitative experimental data to compare with. Thus, the main emphasis has been on comparisons with QM data, although accurate representation of solvent is still difficult. When the requirement was to maintain the experimental native protein conformations within a relatively short simulation time, this was generally considered acceptable. However, the present requirement has raised to the level where the ability to model both the native and nonnative protein energy surface correctly and do so in a realistic solvent environment is needed. Given the limited accuracy of QM data and the lack of realism in solvent representation (or protein interior) in QM calculations, it suffices to say that QM data should be treated only as guidelines for fitting, not the final criteria for judging the accuracy of the force field. We are now, again, confronted with the lack of reliable and quantitative data to compare our results against. One alternative approach is to compare the simulation against data obtained from statistical analysis of high-resolution experimental (protein) structures. Recently, Hu and Hermans[44] studied the energetics of small peptides with a combination of QM/MM approaches in a realistic solvent environment and compared the energy profiles against the PMF data obtained from analysis of high-resolution crystal structures.[45] Their comparisons clearly demonstrated that all of the present force fields have certain degree of bias when compared to the high-level QM/MM data. Indeed, we realize that such comparison is still qualitative in nature, and it should be combined with other data to evaluate the accuracy of the force field. In particular, because dipeptides can not form intramolecular hydrogen bonds, which are present in protein structures, the relative strength of the main chain hydrogen bonds cannot be examined from such tests. Nevertheless, it would be equally erroneous if we completely dismiss such comparison.

Shown in Figure 4 is the Ramanchandran plot obtained from simulations of Ace-Ala-Nme peptide in TIP3P water. Overall, it closely resembles the one obtained by Lovell et al.[45] The main difference between this and the one obtained from statistical analysis of protein structures is the lack of representation in the α_L region, contrary to the conclusion drawn from the comparisons with QM data. Interestingly, the same region was better sampled in the Ace-Ala_4-Nme penta-peptide (discussed later). A striking feature is that the balance between the extended and the helical regions are preserved very well and the left-hand side of this Ramanchandran plot appears to be similar to that of Lovell et al.[45] This is remarkable, given the significance of the balance between the extended and helical regions in protein structure modeling. Thus, we anticipate that this force field would give an improved realism in secondary structure modeling, which has so far been a challenge to the force field development community.

Simulation of Ace-Gly-Nme in TIP3P Water

One of the curious findings in the work of Hu and Hermans[44] is that almost all present force fields match poorly to the main features of the PMF of Ace-Gly-Nme, notwithstanding its small size and simple structure, except for the Cornell et al. force field. With only one hydrogen as the R group, Glycine is the most flexible amino acid. The main features of its PMF surface, as observed from the crystal structures (shown as contours in Fig. 5), include unfavorable regions around $\Phi = 0°$ and around $\Psi = \pm 90°$. From Figure 5, one can see that these features are clearly well represented in the new force field. This is quite possibly the only force field available that has achieved this level of accuracy. Remarkably, the distribution appears to be in better agreement with experiment than with high-level QM/MM simulations by Hu and Hermans.[44] This suggests that inclusion of higher order terms is not the only route to achieve decent accuracy for this type of peptides. Conversely, even with the inclusion of higher order

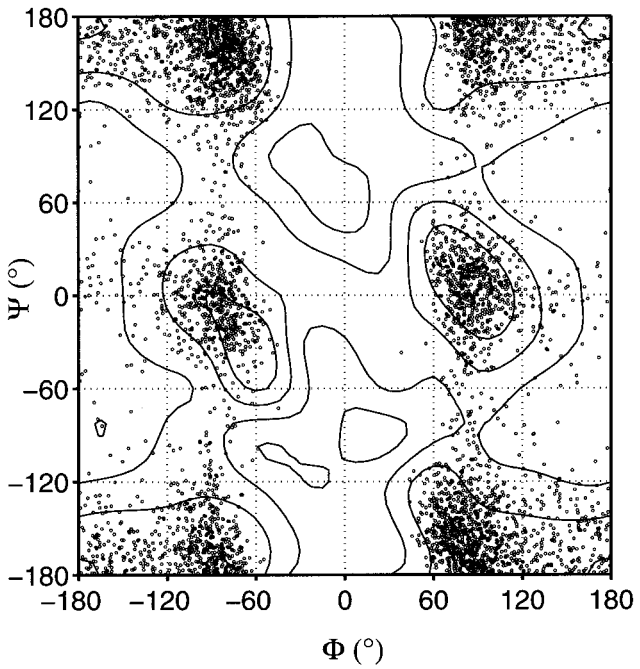

Figure 5. Ramanchandran plot of Ace-Gly-Nme di-peptide from simulation in TIP3P water and is compared to the PMF contours obtained from statistical analysis of high-resolution X-ray structures. Please see refs. 44 and 45 for comparison with other force fields.

terms, as far as the main-chain distribution is concerned, fine tuning of the main-chain torsion parameters will be crucial to achieve good balance between important conformations because of the inevitable approximation. Given the frequent usage of Gly to form turns in peptide design, the new parameter set will facilitate the modeling and design of these peptides. The latter, in turn, will help to assess the accuracy of the model.

Simulation of Ace-Ala$_4$-Nme in TIP3P Water

The Ace-Ala$_4$-Nme peptide is one of the smallest peptides that are capable of forming main-chain hydrogen bonds. In this case, it can potentially form two main-chain hydrogen bonds. Experimental studies on Alanine-based peptides suggest that short peptides up to seven alanine residues are unstructured in solution.[46] This is advantageous for its lack of significant (free) energy traps that reduce the sampling efficiency. In comparison, a structured peptide would impose conformational preference, which is undesirable if we intend to compare with the experimental data obtained from the statistical analysis of a large number of high resolution crystal structures. Shown in Figures 6 are the scattered and PMF contour plots of the main-chain Φ–Ψ distribution of the Ala residues. One of the notable features of the scattered plots is the sampling of the α_L region. This is in contrast to the lack of sampling in the same region in the Ace-Ala-Nme dipeptide simulation. In our opinion, this can be attributed to the interactions of the side chains whose inclusion inevitably change the overall energy surface. Thus, calibration based on small molecules could be potentially misleading.

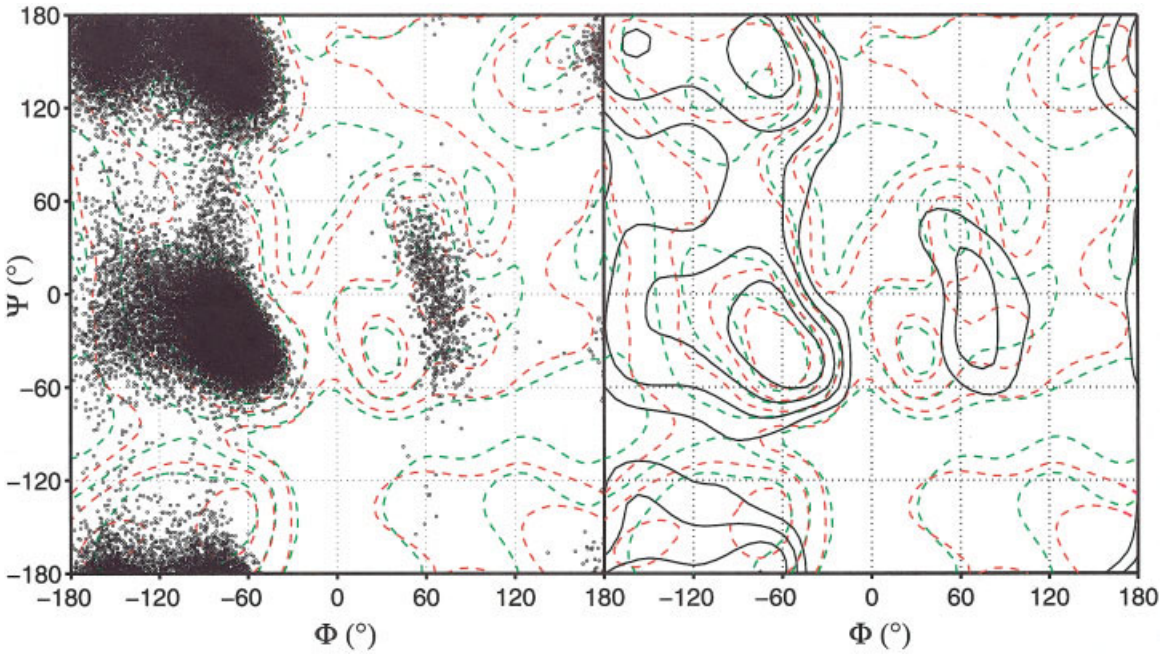

Figure 6. Ramanchandran plot of the Ace-(Ala)$_4$–Nme penta-peptide from simulation in TIP3P water. PMFs from statistical analysis of high-resolution X-ray structures are shown in contours (red and green). Here, the Φ–Ψ distribution obtained from the simulation is shown as both scatters (A) and contours (B). The red dashed contours are Ala residues of from the high-resolution X-ray protein structures, and the green dashed contours are for all residues except Gly and Pro.

Again, the type II poly-proline helical region is clearly the most favorable region, which should be the case for unstructured short peptides as shown by recent experiments.[47]

Comparisons with the PMF contours obtained from statistical analysis of high-resolution X-ray structures further demonstrated that the new force field has achieved a reasonable level of accuracy. Here, we can clearly see that the new force field reproduced important features of the experimental PMF contours, including a reasonable balance between the extended β-sheet conformation and the helical conformation.

The difference found at the type II poly-proline helical region is largely due to the unstructured peptide conformation; this is in agreement with recent experiments, which have shown that the alanine-based short peptides have the tendency to form poly-proline type II helices.[47] The ability to reproduce such a fine feature will greatly enhance the realism in simulations.

Tests on Other Peptides and Proteins

We have conducted extensive tests to examine the accuracy of the force field using a variety of peptide and protein systems, including alanine-based helical peptide,[48] β-sheet–forming peptides (both β-hairpins and three-stranded β-sheets, manuscript in preparation), Trp-cage miniprotein,[11] and a large ensemble of protein decoys.[49] The details of these test results will appear elsewhere. Here, we summarize some of the highlights.

In the alanine-based peptide, a total of 32 simulations[48] were done, and each of the simulations were run to 100 ns, using a Generalized Born continuum solvent model.[7] The calculated average helicity is in excellent agreement with experiments judged by both main-chain hydrogen bonds and main-chain torsion angles.[48] In comparison to simulations on similar peptides using other force fields, there is a noticeable difference in the relative population of three helical species. The equilibrium populations in the π-helical and 3_{10} conformations are, respectively, 4.8 and 1.3%, as measured by the main-chain hydrogen bonds. Both are substantially lower than those in simulations using the CHARMM force field[50,51] (7% for π-helices and 6.5% for 3_{10}-helices). Feig et al. recently studied the three helical species and found that the overrepresentation of the π-helices in the CHARMM force field was likely a force field artifact.[52] The 3_{10}-helical conformation is also much less populated in comparison to simulations[53] using the Cornell et al. force field. In fact, based on our simulations,[48] we believe that both π-helix and 3_{10}-helix are transient species formed by thermal fluctuations from the α-helical conformation as judged by the observation that 61% of the π-helix and 3_{10}-helix species form bifurcated hydrogen bonds. This agrees with a recent study by Armen et al., who compared their simulation results with NMR experimental data.[54] We also observed that two-turn short helices tend to be unstable, which agreed with the experimental observation that seven-amino acid peptides cannot form stable helices.[46,47]

In the β-sheet peptides, simulations (manuscript in preparation) were done on four β-hairpin peptides using a Generalized Born solvent model.[7] About 20 simulations were conducted for each peptide, and the number of trajectories that have reached the β-hairpin conformation ranged from 4 to 13. The extrapolated folding times, based on two-state assumption, ranged from 210 to 440 ns, which are reasonable given that the continuum solvent models accelerate the process by neglecting solvent viscosity. We also successfully folded a three-stranded β-sheet peptide using the same simulation method.

In the case of the Trp-cage miniprotein, a simulation was conducted to 100 ns using a Generalize Born solvent model.[11] The Trp-cage was able to adopt its native structure with the main-chain RMSD of 1.0 Å and heavy-atom RMSD of 2.0 Å from the native NMR structure within 30 ns, and remained in that state until the end of the simulation. In comparison, the heavy-atom RMSDs of the 38 NMR structures range from 1.6 to 2.8 Å. Thus, the simulated structure clearly approached the accuracy of the NMR data. When compared to the simulations using Cornell et al. charge set with a set of tuned main-chain torsion parameters (tuned against a large pool of misfolded peptides[8]) and using an OPLS-AA force field,[10] the simulation using our new parameter set stood out as the best result observed so far. In the work of Simmerling et al.,[8] the simulation successfully folded the Trp-cage protein to a main-chain RMSD between 1.0 to 2.0 Å. In the work of Snow et al.,[10] they studied the folding rates of Trp-cage by approximately 7000 simulations, with the durations ranged from 1.0 to 80.0 ns. Among these massive number of trajectories, only one trajectory reached main-chain RMSD of 1.4 Å with nonnative packing of the Trp side chain (the Trp is flipped by about 90° compare to NMR structure). Incidentally, Generalized Born solvent models were used in all three simulations. Thus, the differences in these simulations are more likely due to the underlying force fields used in the simulations, although details of how the solvent models were implemented may also play a role. In our simulation, it is noteworthy that the Trp side chain was snuggly packed into the cage and remained in that conformation for about 70% of the time.

Further tests were done on a large set of protein decoys.[49] In these tests, a Generalized Born solvent model similar to other studies was used.[7] However, unlike in other similar tests conducted using other force fields[55,56] where the native and decoys were only subjected to energy minimization, we conducted short (10-ps) molecular dynamics simulations on the decoys and used the average energy obtained from the simulations to calculate the *Z*-scores, a measure of the discriminatory ability of the approach. The details of this study have been reported elsewhere. We present here a brief comparison with other similar studies using other force fields[55,56] in Table 5. Clearly, the new approach, the combination of improved protocol (molecular dynamics vs. energy minimization), and the new force field, have significantly improved upon the reported results. In particular, our average *Z*-scores were about 1.0 better than those published by others, which means that our approach has a better discriminatory ability than the previously published results by more than one standard deviation of the energy distribution. Such an enhanced ability is also reflected in the improved fidelity of finding the correct protein structures from the large pool of decoys. Thus, judging from the improved *Z*-scores and the improved ability to differentiate the decoys from the native protein structures, we concluded that this new force field performed better than the other two existing force fields.

Discussion

In molecular mechanics models, torsion potentials serve the role to account for those higher order terms that are otherwise not present

Table 5. Comparison with Other Studies Based on CHARMM 19 and OPLS-AA Force Fields on Protein Decoy Sets.

Decoy Sets	CHARMM/GB[56]	OPLS-AA/GB[55]	AMBER/GB-MD
Four-state Reduced[a]			
% Accuracy	~100%	43%	100%
Average Z_{native}	−3.39	−3.67	−4.95
Average Z'	n/a	n/a	−1.79
Range of Z_{native}	−1.7 to −4.6	−2.18 to −4.53	−2.86 to −6.33
LMDS[b]			
% Accuracy	n/a	14%	80%
Average Z_{native}	n/a	−2.57	−4.49
Average Z'	n/a	n/a	−3.18
Range of Z_{native}	n/a	0.6 to −14.57	4.99 to −8.26
Lu and Skolnick[c]			
% Accuracy	n/a	n/a	93%
Average Z_{native}	n/a	−4.02	−5.82
Average Z'	n/a	n/a	−3.25
Range of Z_{native}	n/a	−1.48 to −9.6	0.34 to −11.63

The "AMBER/GB-MD" column is the data obtained using the new AMBER force field described in this article.

[a,b]Both the "four-state reduced" and the local minima decoy sets (LMDS) were obtained from the Decoys 'R' Us database.[61]

[c]The "Lu and Skolnick" decoy set (courtesy of Lu and Skolnick) contains 54 unique protein sequences, 28 of which has NMR determined structures, and 26 of which has X-ray structures with resolutions ranging from 2.5 to 1.2 Å. The decoys in this set were generated on a lattice using an *ab initio* Monte Carlo structure prediction program,[62,63] and each of these sequences has 1333 decoy structures. "% Accuracy" specifies the percentage of proteins having the lowest energies in comparison to the decoys. "Average Z_{native}" is the average Z-score of the native structure and is calculated as $Z_{native} = (E_{native} - \langle E \rangle_{all})/\sigma$, where E_{native} is the average energy obtained from the MD simulation on the native structure and $\langle E \rangle_{all}$ is that averaged over all structures, including decoys; σ is the standard deviation of the average energies. "Average Z'" is the average Z-score of the native structure in comparison to the decoy of the lowest energy. It is calculated as $Z' = (E_{native} - E_{decoy})/\sigma$, where E_{decoy} is the average energy of the decoy that has the lowest energy among the decoys. The "CHARMM/GB" data were taken from Dominy and Brooks,[56] and the "OPLS-AA/GB" data were from Felts et al.[55]

explicitly in the model. Historically, this has been the most difficult part of the force field development, and will perhaps remain so in the foreseeable future. In particular, the main-chain torsion potentials must reach an extremely high level of accuracy if the force field is intended to model poly-peptide. One of the most crucial issues in achieving this high accuracy is the balance between the helical and the extended conformations. For example, a small bias of 0.1 kcal/mol per residue towards one way or the other would grow to 1.0 kcal/mol when the force field is applied to model a small 10-residue peptide. Obviously, such a level of accuracy is yet unattainable with the present technology, and will perhaps remain a rather challenging task for some time. The problem is further compounded with the requirement of accurate condensed-phase QM data. This is likely true even for a fully polarizable force field because of the truncation of higher order terms. Thus, comparison with QM data is just one of many steps needed to achieve a well-balanced force field.

Traditionally, comparisons with either experimental or high-level quantum mechanical energies of small molecules were taken as important tests for force fields. Such an approach is clearly valuable, and can provide quantitative assessment on the accuracy of the force fields in the confined application areas of those respective small molecular systems. Because our objective is to develop a force field for the simulations of proteins, it is more relevant to test the force field against experimentally well-characterized peptide systems. Our test results clearly showed that the new force field has achieved a reasonable balance in helical and extended conformations. Obviously, more tests have to be done to more fully characterize its behavior and its ability to model proteins. One crucial test would be in the area of tertiary and side-chain contacts. As we stated before, because the charges were derived in condensed phase and may mimic the polarization effect, we are optimistic that the force field will give reasonable performance in this area as well.

A typical approach in the torsion parameter development in the past was to fit the torsion parameters against a few important energy points calculated at relatively high-level quantum mechanical theory (often in gas phase). In our present work, the main-chain torsion parameters were obtained by fitting to a 2D (Φ–Ψ) 144-point energy grid (12 × 12) of alanine-dipeptide calculated using MP2/cc-pVTZ//HF/6-31G** quantum mechanical methods

in organic solvent ($\varepsilon = 4.0$). This is similar to the approach used in the recent refinement of OPLS-AA force field parameters by Kaminski et al.[38] In their approach, the torsion parameters, including those of the side chains, were fitted against gas-phase QM energies. However, the charge set in OPLS-AA force field[57] was tuned against condensed-phase experimental data. As we suggested earlier, such a fitting procedure can be a source of possible overcompensation. The differences in the energy profiles due to environment are also of concern.

In our approach, because the same solvent environment was applied to derive charges and to fit the torsion parameters, overcompensation should be less of a problem. It also ensures that the energetic surface represents the condensed phase. The elaborate map of the entire 2D Φ–Ψ energy surface makes it possible to fit the torsion potentials rationally and ensures that the resulting molecular mechanic energy surface represent the QM surface.

The choice of $\varepsilon = 4.0$ in organic solvent, instead of $\varepsilon = 78.4$ in water, reflects our intention to mimic protein interior. Studies have indicated that the dielectric constant of protein interior is in the neighborhood of $\varepsilon = 10$–12.[58,59] However, measurements on dry protein powder indicated that the dielectric constant, in the absence of buried water (exactly the kind of environment that our model intends to mimic), is between $\varepsilon = 2$ and $\varepsilon = 4$.[60] Thus, the elevated dielectric constant ($\varepsilon = 10$–12 from $\varepsilon = 4.0$) is largely attributed to the presence of water molecules and perhaps also the dynamics of proteins in solution.[58,59] Because this set of parameters are designed for solution phase simulations, solvent effect will be treated either by explicit water or continuum solvent models, which will take into account the dielectric effect of water. Thus, the choice of $\varepsilon = 4.0$ is appropriate. Furthermore, our conservative choice of $\varepsilon = 4.0$ is intended to avoid over polarization. This is important, because overpolarization would exaggerate the charges further and the resulting peptide models would be too hydrophilic. Even though studies have indicated that the IEFPCM model has reached a respectable level of accuracy, it is nevertheless an empirical model. Thus, the choice of $\varepsilon = 4.0$, instead of $\varepsilon = 10$, would, hopefully, give us an additional margin. It is also recognized that the solvent effect, in the absence of explicit solvent molecules, accounts for the free energy difference in respect to the gas phase. The latter is proportional to $(1 - 1/\varepsilon)$. The solvent polarization effect is to minimize this term. One, therefore, intuitively expects that such an effect would also be proportional to $(1 - 1/\varepsilon)$. As a consequence, a medium of $\varepsilon = 4.0$ would produce approximately 75% of polarization effect of a perfect conductor ($\varepsilon = \infty$) when other factors are equal.

The overall agreement between the dipole moments of the new charge set and those of Cornell et al. charges suggests that they are compatible. This implies that the new charge set can be applied directly to study ligand binding with the existing ligand charges, although improvement is likely if these charges can be rederived within the framework of the new charge set. This "backward compatibility" is beneficial, given the significant investment required in the parameterization of organic molecules. It allows a smooth transition to the new frame work.

Conclusion

We presented a novel approach to derive a point-charge model for simulations of proteins in the condensed phase. The charges were based on high-level quantum mechanical electrostatic potentials calculated using continuum solvent model with $\varepsilon = 4.0$. The main-chain torsion parameters of peptides were obtained by fitting to the MP2/cc-pVTZ//HF/6-31G** energy profiles of Alanine and Glycine di-peptides, which were also calculated in $\varepsilon = 4.0$ organic solvent to ensure a balanced parameter set. Our initial test results were encouraging, and clearly showed excellent balance between the extended and helical regions.

Acknowledgment

Computer time was provided by Pittsburgh Supercomputer Center. Helpful suggestions by the reviewers are gratefully acknowledged.

References

1. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T. A.; Lee, H.; Pedersen, L. G. J Chem Phys 1995, 103, 8577.
2. Simmerling, C. L.; Elber, R. Proc Natl Acad Sci USA 1995, 92, 3190.
3. Simmerling, C.; Miller, J. L.; Kollman, P. A. J Am Chem Soc 1998, 120, 7149.
4. Duan, Y.; Kollman, P. A. Science 1998, 282, 740.
5. Crowley, M. F.; Darden, T. A.; Cheatham, T. E.; Deerfield, D. W. J Supercomp 1997, 11, 255.
6. Bashford, D.; Case, D. A. Annu Rev Phys Chem 2000, 51, 129.
7. Tsui, V.; Case, D. A. J Am Chem Soc 2000, 122, 2489.
8. Simmerling, C.; Strockbine, B.; Roitberg, A. E. J Am Chem Soc 2002, 124, 11258.
9. Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. Nature 2002, 420, 102.
10. Snow, C. D.; Zagrovic, B.; Pande, V. S. J Am Chem Soc 2002, 124, 14548.
11. Chowdhury, S.; Lee, M. C.; Xiong, G.; Duan, Y. J Mol Biol 2003, 327, 711.
12. Shakhnovich, E. I. Curr Opin Struct Biol 1997, 7, 29.
13. Sheinerman, F. B.; Brooks, C. L. Proc Natl Acad Sci USA 1998, 95, 1562.
14. Duan, Y.; Wang, L.; Kollman, P. A. Proc Natl Acad Sci USA 1998, 95, 9897.
15. Voter, A. F. Phys Rev B 1998, 57, 985.
16. Zagrovic, B.; Sorin, E. J.; Pande, V. S. J Mol Biol 2001, 313, 151.
17. Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. J Comp Chem 1986, 7, 230.
18. Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. Science 1987, 236, 564.
19. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
20. Cheatham, T. E., III; Kollman, P. A. J Mol Biol 1996, 259, 434.
21. Duan, Y.; Patricia, W.; Crowley, M.; Rosenberg, J. M. J Mol Biol 1997, 272, 553.
22. Cieplak, P.; Caldwell, J.; Kollman, P. A., personal communication.
23. Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A. J Comp Chem 2002, 23, 1515.

24. Wang, J. M.; Kollman, P. A. J Comp Chem 2001, 22, 1219.
25. Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. J Phys Chem 1993, 97, 10269.
26. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; J. R. Cheeseman, V. G. Z.; J. A. Montgomery, J.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; 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.; Andres, J. L.; Gonzalez, C.; Head–Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98; Gaussian, Inc.: Pittsburgh, PA, 2001.
27. Lee, C.; Yang, W.; Parr, R. G. Phys Rev B 1988, 37, 785.
28. Becke, A. D. J Chem Phys 1993, 98, 5648.
29. Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Chem Phys Lett 1989, 157, 200.
30. Kendall, R. A.; Dunning, T. H.; Harrison, R. J. J Chem Phys 1992, 96, 6796.
31. Tomasi, J.; Mennucci, B.; Cances, E. Theochem-J Mol Struct 1999, 464, 211.
32. Pomelli, C. S.; Tomasi, J.; Barone, V. Theor Chem Acc 2001, 105, 446.
33. Cances, E.; Mennucci, B.; Tomasi, J. J Chem Phys 1997, 107, 3032.
34. Kollman, P. A.; Dixon, R. W.; Cornell, W. D.; Fox, T.; Chipot, C.; Pohorille, A. In Computer Simulations of Biological Systems; van Gunsteren, W. F., Ed.; Escom: The Netherlands, 1997.
35. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. J Chem Phys 1983, 79, 926.
36. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. J Comp Phys 1984, 81, 3684.
37. Wang, G.; Dunbrack, R. L. J., Bioinformatics, 2003, 9, 1589.
38. Kaminski, G. A.; Friesner, R. A.; Tirado–Rives, J.; Jorgensen, W. L. J Phys Chem B 2001, 105, 6474.
39. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. J Am Chem Soc 1997, 119, 5908.
40. Klamt, A.; Schuurmann, G. J Chem Soc Perkin Trans 1993, 2, 799.
41. Klamt, A. J Phys Chem 1995, 99, 2224.
42. Andzelm, J.; Kolmel, C.; Klamt, A. J Chem Phys 1995, 103, 9312.
43. Klamt, A.; Jonas, V. J Chem Phys 1996, 105, 9972.
44. Hu, H.; Hermans, J. Proteins 2003, 50, 451.
45. Lovell, S. C.; Davis, I. W.; Arendall, W. B. III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Proteins 2003, 50, 437.
46. Rohl, C. A.; Baldwin, R. L. Biochemistry 1997, 36, 8435.
47. Shi, Z. S.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. Proc Natl Acad Sci USA 2002, 99, 9190.
48. Chowdhury, S.; Zhang, W.; Wu, C.; Xiong, G.; Duan, Y. Biopolymers 2003, 68, 63.
49. Lee, M. C.; Duan, Y. Proteins 2003, in press.
50. Ferrara, P.; Apostolakis, J.; Caflisch, A. J Phys Chem B 2000, 104, 5000.
51. Shirley, W. A.; Brooks, C. L., III. Proteins 1997, 28, 59.
52. Feig, M.; MacKerell, A. D.; Brooks, C. L. J Phys Chem B 2003, 107, 2831.
53. Sung, S. S.; Wu, X. W. Proteins 1996, 25, 202.
54. Armen, R.; Alonso, D. O. V.; Daggett, V. Protein Sci 2003, 12, 1145.
55. Felts, A. K.; Gallicchio, E.; Wallqvist, A.; Levy, R. M. Proteins 2002, 48, 404.
56. Dominy, B. N.; Brooks, C. L. J Comp Chem 2002, 23, 147.
57. Jorgensen, W. L.; Maxwell, D. S.; Tirado–Rives, J. J Am Chem Soc 1996, 118, 11225.
58. GarciaMoreno, B.; Dwyer, J. J.; Gittis, A. G.; Lattman, E. E.; Spencer, D. S.; Stites, W. E. Biophys Chem 1997, 64, 211.
59. Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; Garcia–Moreno, B. Biophys J 2000, 79, 1610.
60. Harvey, S. C.; Hoekstra, P. J Phys Chem 1972, 76, 2987.
61. Park, B.; Levitt, M. J Mol Biol 1996, 258, 367.
62. Kihara, D.; Lu, H.; Kolinski, A.; Skolnick, J. Proc Natl Acad Sci USA 2001, 98, 10125.
63. Lu, H.; Skolnick, J. Proteins 2001, 44, 223.