JCTC Journal of Chemical Theory and Computation

Effective Approximation of Molecular Volume Using Atom-Centered Dielectric Functions in Generalized Born Models

Jianhan Chen*

Department of Biochemistry, Kansas State University, Manhattan, Kansas 66506

Received May 12, 2010

Abstract: The generalized Born (GB) theory is a prime choice for implicit treatment of solvent that provides a favorable balance between efficiency and accuracy for reliable simulation of protein conformational equilibria. In GB, the dielectric boundary is a key physical property that needs to be properly described. While it is widely accepted that the molecular surface (MS) should provide the most physical description, most existing GB models are based on van der Waals (vdW)-like surfaces for computational simplicity and efficiency. A simple and effective approximation to molecular volume is explored here using atom-centered dielectric functions within the context of a generalized Born model with simple switching (GBSW). The new model, termed GBSW/MS2, is as efficient as the original vdW-like-surface-based GBSW model, but is able to reproduce the Born radii calculated from the "exact" Poisson−Boltzmann theory with a correlation of 0.95. More importantly, examination of the potentials of mean force of hydrogen-bonding and charge−charge interactions demonstrates that GBSW/MS2 correctly captures the first desolvation peaks, a key signature of true MS. Physical parameters including atomic input radii and peptide backbone torsion were subsequently optimized on the basis of solvation free energies of model compounds, potentials of mean force of their interactions, and conformational equilibria of a set of helical and β-hairpin model peptides. The resulting GBSW/MS2 protein force field reasonably recapitulates the structures and stabilities of these model peptides. Several remaining limitations and possible future developments are also discussed.

1. Introduction

An accurate description of the solvent environment is critical in molecular modeling of biomolecule structure and dynamics. Traditional explicit inclusion of water molecules arguably provides the most detailed description of the solvent but, at the same time, dramatically increases the system size and thus the associated computational cost. The expensive computational cost further hinders extensive optimization of explicit solvent force fields for accurate description of protein conformational equilibria.[1,2] Instead, implicit treatment of the solvent environment has recently emerged as a powerful alternative to explicit water in biomolecular modeling.[3] Implicit solvent essentially aims to capture the mean influence of solvent molecules on the solute through direct estimation of the solvation free energy, defined as the reversible work required to transfer the solute from vacuum to solution in a fixed configuration. Elimination of the solvent molecules with the implicit treatment substantially reduces the number of atoms needed to be simulated. More importantly, this can now be achieved with only a moderate increase in the computational overhead required for estimating the solvation free energy on-the-fly, such as using continuum electrostatics-based methods including Poisson−Boltzmann (PB) and generalized Born (GB) theories.[4−6] Such a dramatic reduction in the system size does not come without a loss of detail and certain intrinsic limitations. In general, implicit solvent models may yield considerable disagreement with explicit water simulations in short-range effects when the detailed interplay of a few water molecules

* Phone: (785) 532-2518; fax: (785) 532-7278; e-mail: jianhanc@ksu.edu.

10.1021/ct100251y
Published on Web 08/20/2010

(which are distinct from the bulk water) is important.[7,8] Such limitations have motivated several attempts to better describe the short-range effects either through empirical corrections[9] or using hybrid explicit/implicit representations.[10,11] Implicit solvent might be further limited by the specific methodology for estimating the solvation free energy, as well as the (physical) parameters of the solvation model and underlying protein force field. Nonetheless, the substantial reduction in the computational cost and extension of accessible simulation time scales with implicit treatment of solvent is an important advantage. It not only allows careful optimization of the force field through extensive peptide simulations,[12−15] but has also opened a door to address many biological problems that are otherwise difficult with explicit solvent.[4,6]

Among various approaches for implicit treatment of the solvent, GB has been particularly successful for molecular dynamics simulation of biomolecules and especially proteins.[6] In GB, the total solvation free energy is decomposed into nonpolar and electrostatic contributions:[3]

$$\Delta G_{\mathrm{solv}} = \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{np}} \tag{1}$$

Such a decomposition is path-dependent, but it allows both contributions to be related to appropriate (continuum) models of water and is generally more accurate than fully empirical approaches that directly estimate the total solvation free energy from certain solute geometric properties.[16,17] Given the well-established continuum electrostatics description of water, where the solute is represented as a low dielectric cavity embedded in a featureless, high dielectric solvent medium, the electrostatic solvation free energy can be rigorously calculated by solving the PB equation using finite-difference methods.[18−20] Alternatively, the GB pairwise approximation can be used to calculate the same quantity:[21,22]

$$\Delta G_{\mathrm{elec}} = -\frac{1}{2}\tau \sum_{ij} \frac{q_i q_j}{\sqrt{r_{ij}^2 + R_i^{\mathrm{GB}} R_j^{\mathrm{GB}} \exp(-r_{ij}^2/F R_i^{\mathrm{GB}} R_j^{\mathrm{GB}})}} \tag{2}$$

where r_{ij} is the distance between atoms i and j and q_i is the atomic charge and R_i^{GB} the *effective Born radius* of atom i. $\tau = 1 - 1/\varepsilon_s$, with ε_s being the solvent (high) dielectric constant. F is an empirical factor whose value may range from 2 to 10, with 4 being the most common one. The GB approximation is much more efficient than PB computationally. At the same time, it provides analytical forces and is thus particularly suitable for molecular dynamic simulations. Accurate calculation of the nonpolar solvation free energy is much more challenging for biomolecules with complex shapes. At present, the nonpolar solvation free energy is either largely ignored or estimated directly from the solvent-accessible surface area (SA) with a phenomenological surface tension coefficient, $\Delta G_{\mathrm{np}} = \gamma S$. With substantial improvement in the electrostatic solvation models, limitations of simple SA models are becoming increasingly important for accurate simulation of protein conformational equilibria.[23−25]

The effective Born radius, R_i^{GB}, is a key quantity in the GB formalism. It corresponds to the distance between a particular atom and its hypothetical spherical dielectric boundary, chosen such that the atomic electrostatic self-solvation energy satisfies the Born equation[26]

$$\Delta G_{\mathrm{elec},i} = -\frac{1}{2}\tau \frac{q_i^2}{R_i^{\mathrm{GB}}} \tag{3}$$

In principle, the "exact" effective Born radii can be calculated from eq 3 using the electrostatic self-solvation energy obtained through the PB theory. It has been shown that, given accurate effective Born radii, the GB approximation of eq 2 closely reproduces the PB electrostatic solvation energy.[27,28] As such, most of the extensive literature on extensions of the GB theory has been focused on efficient and accurate evaluation of the effective Born radii, and PB-derived $\Delta G_{\mathrm{elec},i}$ or R_i^{GB} have served as standard benchmarks for assessing the numerical accuracy of various GB approximations.[9,29−41] At present, the GB formalism has reached a mature stage, and many of the latest models are capable of achieving a level of numerical accuracy that approaches that of high-resolution PB calculations.[28] Such numerical accuracy provides a necessary basis for recent efforts to optimize various GB-based implicit solvent protein force fields for an accurate description of peptide and protein conformational equilibria.[12−15]

It should be emphasized that most GB models achieve high numerical accuracy through careful optimization of what can be referred to as "numerical parameters": parameters specific to a particular GB formalism that are adjusted to maximally reproduce the exact results from equivalent high-resolution PB calculations. Key quantities that need to be reproduced in reference to PB include solvation free energies of small model compounds as well as proteins with various sizes and structures and, most importantly, the effective Born radii (see eq 3) in complex protein environments. These numerical parameters should be distinguished from "physical parameters", such as the definition of the dielectric boundary (if adjustable), the intrinsic atomic radii for defining the location of the dielectric boundary, and parameters associated with the nonpolar solvation component (e.g., effective surface tension coefficients). These parameters all have well-defined physical meanings and should be optimized to reproduce certain (experimental) physical properties. In particular, optimization of the physical parameters is meaningful when and only when satisfactory numerical accuracy has been achieved. Otherwise, improper cancellation of errors might occur, and this limits the transferability of the optimized implicit solvent force field. It should be noted that there has been some confusion in the literature concerning the optimization of GB implicit solvent, where numerical accuracy is often confused with physical accuracy or physical accuracy is discussed without first establishing the numerical accuracy of the method. Importantly, careful optimization of physical parameters is also necessary for reliable application of various PB methods to biomolecular modeling.

We have previously optimized the intrinsic atomic radii and backbone torsion potentials of the GBSW (generalized Born with simple switching) implicit solvent[38] together with the underlying CHARMM param22/CMAP protein force field[42−45] for reliable simulation of peptide conformational

equilibrium.[12,13] The optimized force field appears to achieve a reasonable balance of competing solvation and intermolecular interactions and successfully folds a set of diverse model peptides and miniproteins. It has also been applied to model the conformational equilibria of stable and unstable states of several small proteins,[46−48] including two intrinsically disordered proteins.[49,50] Nonetheless, application of this force field to folding of larger proteins has not been as satisfactory, and the simplistic SA-based treatment of nonpolar solvation has been proposed to be a key limitation.[25]

The focus of the current work is to explore and address another methodological limitation of GBSW in the description of the solvent−solute dielectric boundary. Specifically, GBSW utilizes an atomic-centered function to smoothly switch between high (water) and low (solute) dielectric regions.[38,51] Such a smooth dielectric function captures a van der Waals (vdW)-like surface and is computationally efficient and numerically stable. Similar atom-centered dielectric functions have also been widely utilized in many GB (and some PB) models.[6,28] However, it has been recognized that vdW-like surface definitions result in small, solvent-inaccessible (and thus unphysical) high dielectric pockets, which can lead to significant overestimation of the solvation free energy for larger proteins.[10,52,53] Instead, the Lee−Richards molecular surface (MS),[54] defined by rolling a (solvent) probe sphere over the surface of the solute molecule, arguably provides the most appropriate dielectric boundary. Adopting MS in implicit solvent models is very challenging, though, due to a lack of analytical definition, discontinuities with respect to infinitesimal atomic displacements, and sensitivity to grid discretization.[37] Several attempts have been made in approximating the original Lee−Richards MS using analytical functions in the context of GB during recent years with various levels of success.[9,55−57] The GBMV2 (generalized Born using molecular volume) model developed by Lee et al.[55] has been one of the most successful models in the ability to reproduce PB-derived effective Born radii and total solvation free energies of proteins. Nonetheless, GBMV2 is substantially more expensive than comparable vdW-like surface-based GB models such as GBSW in computational cost.[28] More importantly, the sharp molecular surface definition leads to unstable atomic forces and poor energy conservation properties.[58] A possible remedy is to adopt a smoother dielectric transition, which reduces the ability to approximate the true MS, and at the same time decreases molecular dynamics (MD) time steps to 1−1.5 fs from 2 fs, which further increases the computational cost.[58] The numerical instability of GBMV2 is an important limitation and has contributed to the difficulty in optimization of the GBMV2 protein force field using similar strategies that prove effective for optimizing GBSW[13] (Chen, unpublished data).

In the rest of this paper, we will first explore the feasibility and effectiveness of approximating the molecular volume purely on the basis of atom-centered dielectric functions within the framework of GBSW. It will be demonstrated that the new model, termed GBSW/MS2, while as efficient and as stable as the original GBSW model, is able to reproduce the effective Born radii calculated from the MS PB theory with a correlation of 0.95. More importantly, potentials of mean force (PMFs) of pairwise polar interactions demonstrate that GBSW/MS2 can correctly capture the first desolvation peak, which is a key signature of a true MS-like surface. We then describe further efforts to extensively optimize the GBSW/MS2 protein force field on the basis of solvation free energies, pairwise interactions of a set of amino acid backbone and side chain analogues, and conformational equilibria of several model peptides. At the end, several important remaining limitations and possible directions for further improvement will be discussed.

2. Methods

2.1. Higher Order Corrections to the Coulomb Field Approximation. The original GBSW model[38] estimates the atomic electrostatic self-solvation free energy on the basis of the Coulomb field approximation (CFA) with a higher order correction term:[37]

$$\Delta G_{\mathrm{elec},i} = a_0 \Delta G^0_{\mathrm{elec},i} + a_1 \Delta G^1_{\mathrm{elec},i} \tag{4}$$

$$\Delta G^0_{\mathrm{elec},i} = -\frac{1}{2}\tau q_i^2 \left(\frac{1}{\eta} - \frac{1}{4\pi} \int_{r>\eta_i} \frac{V(\mathbf{r};\{\mathbf{r}_\alpha\})}{|\mathbf{r}-\mathbf{r}_i|^4}\, \mathrm{d}\mathbf{r} \right) \tag{5}$$

$$\Delta G^1_{\mathrm{elec},i} = -\frac{1}{2}\tau q_i^2 \left(\frac{1}{4\eta^4} - \frac{1}{4\pi} \int_{r>\eta_i} \frac{V(\mathbf{r};\{\mathbf{r}_\alpha\})}{|\mathbf{r}-\mathbf{r}_i|^7}\, \mathrm{d}\mathbf{r} \right)^{1/4} \tag{6}$$

where η is an arbitrarily chosen integration starting point less than or equal to the vdW radius of atom i necessary to avoid the singularity at $r = |\mathbf{r} - \mathbf{r}_i| = 0$. The solute interior volume function, $V(\mathbf{r};\{\mathbf{r}_\alpha\})$, is a function of all atomic positions, $\{\mathbf{r}_\alpha\}$. It is defined by overlapping atom-centered dielectric functions with smooth switching at the solute−solvent boundary:[51]

$$V(\mathbf{r};\{\mathbf{r}_\alpha\}) = 1 - \prod_i H_i(|\mathbf{r}-\mathbf{r}_i|) \tag{7}$$

where the atomic volume exclusion function, $H_i(r)$, is given as

$$H_i(r) = \begin{cases} 0, & r \le R_i - w \\ \frac{1}{2} + \frac{3}{4w}(r-R_i) - \frac{1}{4w^3}(r-R_i)^3, & R_i - w < r < R_i + w \\ 1, & r \ge R_i + w \end{cases} \tag{8}$$

where R_i is the *atomic input radius* to define the solvent−solute dielectric boundary and $2w$ is the smooth switching length. Both $H_i(r)$ and its first derivative are continuous. The value of $V(\mathbf{r};\{\mathbf{r}_\alpha\})$ as defined is 1 in the solute interior and gradually decreases to 0 in the solute exterior. The volume integrals in eqs 5 and 6 are evaluated using an efficient numerical quadrature technique.[37]

The coefficients a_0 and a_1 in eq 4 are two key numerical parameters in GBSW that are parametrized to reproduce the exact $\Delta G_{\mathrm{elec},i}$ computed from PB with the same dielectric boundary definition.[51] The higher order CFA correction term in eq 6 is an empirical one, and the values of $\Delta G^0_{\mathrm{elec},i}$ and $\Delta G^1_{\mathrm{elec},i}$ (or other higher order terms[37,59]) are highly cor-

related. Such empirical corrections to CFA are important for accurately reproducing the corresponding PB results in both GBSW and GBMV. One way to rationalize the effectiveness of empirical expressions such as eq 4 is that one should be able to rewrite the exact atomic self-electrostatic solvation free energy as an expansion that includes the CFA approximation and a series of higher order (correction) terms, even though in practice only one such correction term appears to be sufficient.[37,38,55] In analogy, one might suspect that similar sums of a series of correction terms might also provide a means to approximate MS-like surfaces using atom-centered dielectric functions in a purely empirical fashion.

2.2. Numerical Approximation of the Molecular Surface. In this work, we explore a general expression that is directly parametrized on the basis of MS PB results to approximate an MS-like dielectric boundary:

$$1/R_i^{GB} = D + C_0A_4 + C_1A_7 \quad (9)$$

where $A_4 = -2\Delta G^0_{elec,i}/\tau {q_i}^2$ and $A_7 = -2\Delta G^1_{elec,i}/\tau {q_i}^2$. This expression differs slightly from an analogous one used in GBMV2,[28] in that $1/R_i^{GB}$ instead of R_i^{GB} is used. This is mainly to reflect the importance of reproducing small R_i^{GB}, as they correspond to cases with larger contributions to electrostatic solvation energetics. Parametrization based on $1/R_i^{GB}$ is also more suitable compared to those based on $\Delta G_{elec,i}$ (e.g., eq 4). The reason is that the natural spread of atomic partial charges in proteins superficially increases the correlation between values of $\Delta G_{elec,i}$ from GB and PB (via the factor ${q_i}^2$), which impedes direct optimization of the ability to mimic MS. Implementation of eq 9 requires minimal changes to the original GBSW module[38] available as part of the CHARMM program.[60,61] This new numerical approximation is referred to as GBSW/MS2. Feig et al. also previously parametrized GBSW directly to reproduce MS PB-derived $\Delta G_{elec,i}$ for the case of $w = 0.2$. This optimization was meant as a quick test of mimicking MS within the GBSW framework and has not been extensively tested. It will be referred to as GBSW/MS hereafter. The GBSW/MS fit is equivalent to assigning $D = 0$, $C_0 = 1.204$, and $C_1 = 0.187$ in eq 9. We note that additional higher order terms could be included in eq 9. However, they do not appear to further improve the goodness of fit, likely because the values of these terms are highly correlated. The effectiveness of approximating an MS-like surface will be examined on the basis of the ability to reproduce MS PB results including atomic effective Born radii and total electrostatic solvation free energies of a protein test set. A key signature of the true MS surface is the inclusion of re-entrant surfaces,[62] which are manifested as the first major desolvation peaks in the PMFs of pairwise interactions. The ability to describe the first desolvation peak might serve as an important validation of whether such empirical parametrization can sufficiently mimic an MS-like dielectric boundary. Importantly, the proposed reparametrization does not change the underlying smooth definition of the dielectric boundary, and thus, the new model is as efficient and as stable as the original GBSW.

2.3. Optimization of Physical Parameters. Once the numerical accuracy of GBSW/MS2 was established, key physical parameters of the GBSW/MS2 protein force field such as atomic input radii ($\{R_i\}$), a sufficient tension coefficient (γ), and peptide backbone torsion energetics were optimized on the basis of a set of experimental and theoretical properties. The optimization strategy, briefly summarized below, is analogous to what was utilized previously to optimize the original GBSW protein force field.[13] In principle, other protein force field parameters, particularly Lennard-Jones parameters and atomic partial charges, need to be co-optimized with the new solvation model to achieve full consistency and maximal transferability. However, such an attempt to reparametrize the underlying protein model appears to be highly ambitious. As a compromise and first step, it should be reasonable to focus primarily on directly adjusting the input radii and backbone torsion energetics. An important caveat of such a limited optimization strategy is that one might incorrectly compensate for certain artifacts of the underlying protein model.

2.3.1. Solvation Free Energies and Potentials of Mean Force. The solvation free energies of amino acid side chain analogues are among the few types of experimental data that can be directly used in protein force field parametrization. However, key to an accurate description of peptide conformational equilibria is the ability to capture the delicate balance between sets of competing interactions, i.e., the solvation preference of side chains and backbones versus the strength of (solvent-mediated) interactions between these moieties in a complex protein environment. These two opposing effects are both large and mostly cancel each other. As such, small relative errors in either term might translate into a substantial shift in the balance. Therefore, a more effective approach is to optimize the solvation model directly on the basis of its ability to capture the balance of solvation and intermolecular interactions. Specifically, important physical parameters such as the atomic input radii and surface tension coefficient were systematically optimized to reproduce the strengths of a total of 44 pairwise and three-body interactions among polar and nonpolar model compounds in the TIP3P explicit solvent. A list of all 19 polar PMFs and 25 nonpolar PMFs is provided in the Supporting Information (Figures S1 and S2). Calculation of these TIP3P PMFs has been described in detail in our previous works.[13,24,25] PMFs in implicit solvent were computed by directly translating the molecules along the axis of interaction. Experimental and TIP3P solvation free energies of amino acid side chain analogues[63] were only used for postoptimization validation in this work.

2.3.2. Conformational Equilibria of Model Peptides. After initial optimization of input radii based on PMFs, an iterative procedure was used to empirically fine-tune the solvation parameters together with the peptide backbone torsion energetics, guided by simulation of conformational equilibria of a set of model peptides. The main objective is to balance the solvation model with the underlying CHARMM param22/CMAP protein force field,[42−45] such that both the experimental structures and stabilities of these model peptides can be recapitulated. The model peptides include (1) $(AAQAA)_3$

(~50% helical at 270 K[64]), (2) GB1p (GEWTY DDATK TFTVT E, β-hairpin, 42% folded at 278 K[65]), (3) GB1m1 (GEWTY DDATK TATVT E, β-hairpin, 6% folded at 298 K[66]), (4) HP5A (KKYTW NPATG KATVQ E, β-hairpin, 21% folded at 298 K[66]), and (5) GB1m3 (KKWTY NPATG KFTVQ E, β-hairpin, 86% folded at 298 K[66]). Consistent with the experimental conditions, the termini of the $(AAQAA)_3$ peptide were blocked with Ace and NH_2, and all the other peptides were simulated with unblocked termini. The hairpins GB1m1, HP5A, and GB1m3 are derived from the native sequence of the C-terminal β-hairpin (residues 41−56) of the B1 domain of protein G (GB1p), but display reduced or enhanced stability: (unfolded) GB1m1 < HP5A < GB1p < GB1m3 (most folded).[66] Therefore, these peptide sequences provide a particularly useful control for protein force field optimization.

Heavy reliance on simulation of peptide conformational equilibria is an important aspect of the current optimization strategy. The ability to recapitulate the experimental structures *and* stabilities of the above model peptides provides key feedback for the parametrization of both the atomic input radii and peptide backbone torsion energetics. An important limitation, however, is the slow convergence of conformational equilibria even for small β-hairpins. In this work, we mainly rely on replica exchange molecular dynamics (REMD), as implemented in the MMTSB Toolset[67] (available from http://mmtsb.org), to improve convergence. While REMD has been shown to be able to provide enhanced conformational sampling for cases with positive enthalpies of activation, important questions remain in the true efficiency and optimal setups for sampling protein conformations in current implicit or explicit solvent protein force fields.[68−72] In this work, we chose to adopt REMD setups similar to those used in a previous work,[13] i.e., 16 replicas distributed exponentially within temperatures of 270−500 K. Convergence of simulated conformational ensembles is tested by comparing two independent "control" and "folding" simulations, initially from folded and fully extended conformations, respectively. Such independent simulations from completely different initial coordinates are much more reliable in establishing convergence than simply following the time evolution of the simulations. Additional simulations were also carried out with temperature spans of 270−400 or 270−800 K to seek potential improvement in convergence. Exchanges of simulation temperatures of replicas were attempted every 2 ps, and more frequent exchanges did not appear to improve sampling in any of our tests. The length of the REMD simulation is 20 ns for $(AAQAA)_3$ and ranges from 50 to 150 ns for GB1p series hairpins. The actual exchange acceptance ratios ranged from about 30% to over 70%. As will be discussed in the Results and Discussion, critical limitations appear to exist in obtaining converged conformational equilibria for several hairpins with GBSW/MS2. However, it is not obvious whether this is reflecting a much greater sampling requirement due to the presence of significant desolvation barriers with the underlying MS-like surfaces or one might be able to fine-tune REMD for individual peptides to improve the sampling efficiency.

Given the significant challenges and expensive computational costs of obtaining converged conformational equilibria of even small model peptides, the iterative optimization relies heavily on manual adjustment of the input radii of important backbone atoms (including amide nitrogen and carbonyl oxygen) and peptide backbone dihedral energetics. Modification of the backbone dihedral energetics was realized using the CMAP dihedral cross-term facility in CHARMM.[43−45] As a proper balance of secondary structure preference is one of the primary goals, the modifications were focused on the extended (β) and helical regions of the ϕ/ψ space. Stabilization (or destabilization) of particular conformations was achieved by adding cosine-shaped "valleys" (or "humps") centered at the appropriate ϕ/ψ coordinates to a quantum mechanical CMAP:[43]

$$\Delta E(\phi,\psi) = \frac{1}{2}k^{\alpha}[1+\cos(d^{\alpha}\pi/r^{\alpha})] + \frac{1}{2}k^{\beta}[2+\cos(d_1^{\beta}\pi/r^{\beta}) + \cos(d_2^{\beta}\pi/r^{\beta})] \quad (10)$$

with

$$d^{\alpha} = \min[r^{\alpha}, \sqrt{(\phi-\phi^{\alpha})^2+(\psi-\psi^{\alpha})^2}]$$

and

$$d_l^{\beta} = \min[r^{\beta}, \sqrt{(\phi-\phi_l^{\beta})^2+(\psi-\psi_l^{\beta})^2}] \qquad l=1,2$$

where the centers and radii are $(\phi^{\alpha}, \psi^{\alpha}) = (-75°, -45°)$ with $r^{\alpha} = 60°$ (for the α-helical region) and $(\phi_1^{\beta}, \psi_1^{\beta}) = (-120°, 125°)$ and $(\phi_2^{\beta}, \psi_2^{\beta}) = (-150°, 160°)$ with $r^{\beta} = 45°$ (for parallel and antiparallel β-strand regions. The functional form of eq 10 is purely empirical and mainly serves to reduce the number of adjustable parameters during iterative optimization. It should be noted that the current optimization focuses on describing conformational equilibria, and the strategy of directly targeting local ϕ/ψ regions might have nontrivial consequences on the kinetics of transitions between various conformational states. Another compromise made here is that the details of the unfolded states, such as a prevalence of poly-L-proline II (PPII) helix-like and α_R helix structures,[73] are not explicitly considered.

2.3.3. Control Simulation of Peptides and Proteins. Additional control simulations were also carried out for a few proteins and protein complexes of various sizes and folds to validate the stability of these proteins in the optimized GBSW/MS2 force field. These systems include the B1 domain of protein G (mixed helix/β, PDB 3gb1), the apo form of dihydrofolate reductase (mixed helix/β, PDB 1rx2), and the complex between domains of CREB binding protein (CBP) and the activator of thyroid and retinoid receptors (ACTR) (helical, PDB 1kbh). In addition, the conformational equilibrium of a short polyalanine peptide, Ala_5, is calculated using a 40 ns replica exchange simulation to more directly examine the ability of the optimized GBSW/MS2 force field to describe unfolded protein states. Eight replicas spanning 300−500 K were used, and exchanges of replica temperatures were attempted every 2 ps. Consistent with the experimental conditions (carried out at pH 2), both the N- and C-termini of Ala_5 are protonated. NMR scalar coupling

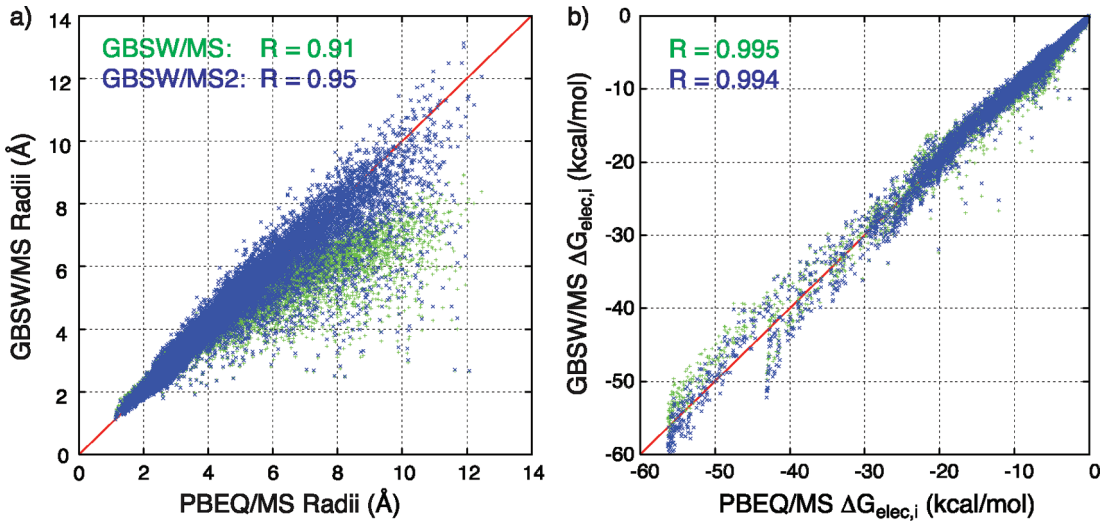

Figure 1. Comparison between PB and GB effective Born radii (a) and atomic electrostatic self-solvation free energies (b) for all atoms of a set of 22 small proteins. Green and blue points correspond to GBSW/MS and GBSW/MS2 results, respectively.

constants are calculated on the basis of the Karplus equation using the same parameters as in the NMR study,[73] specifically ${}^3J(H_N,H_\alpha) = 7.09\cos^2(\phi - 60°) - 1.42\cos(\phi - 60°) + 1.55$ and ${}^2J(N,C_\alpha) = -0.66\cos^2\psi_{i-1} - 1.52\cos\psi_{i-1} + 7.85$.

3. Results and Discussion

3.1. Numerical Parametrization: Effective Born Radii and Total Protein Solvation Free Energies. Numerical parameters in eq 9 were obtained by minimizing the root-mean-square error between GB and PB results for all atoms in a set of 22 small proteins (test set 1 of Feig et al.;[28] also see Table 1 of the Supporting Information). The sizes of these proteins range from 37 to 98 residues. The MS PB results were computed using the PBEQ module[51] in CHARMM with a grid spacing of 0.21 Å and a probe radius of 1.4 Å. The switching length in eq 8 is set to $2w = 0.4$ Å. A previously optimized set of input radii[74,75] (hereinafter referred to as Nina's radii) was used in both GB and PB calculations at this stage. The best fit was achieved with $D = -0.0505$, $C_0 = 1.437$, and $C_1 = 0.1631$, with a correlation coefficient of $R = 0.98$ between GB- and PB-derived $1/R_i^{GB}$ values. Figure 1 compares the PB and GB effective Born radii and electrostatic self-solvation free energies. The correlations of the self-solvation free energies of GBSW/MS and GBSW/MS2 with PB are similar, and both are somewhat misleadingly high, with $R \approx 0.995$ due to the natural spread of atomic partial charges. This reinforces the notion that self-solvation free energies are not sufficiently sensitive for numerical parametrization of GB. The effective Born radii are the most important quantities to be estimated accurately in a GB model. GBSW/MS2 improves the correlation between GB- and PB-derived effective Born radii to $R = 0.95$ compared to $R = 0.91$ for GBSW/MS. While such a correlation is still substantially lower than that of GBMV2 ($R \approx 0.99$), it is a significant improvement over that of $R = 0.811$ achieved by a GSGB model developed by Yu and co-workers.[56] The two other previous attempts to incorporate MS-like surfaces in GB[9,57] do not report numerical values for the correlation between GB and PB effective Born radii. Nonetheless, it might be estimated that the GBn model described by Morgan and co-workers does not perform better than GSGB in reproducing PB-derived effective Born radii (on the basis of Figure 3 of ref 57).

The effectiveness of numerical parametrization in reproducing the total electrostatic solvation free energies was examined using both test set 1 and a larger set of 611 proteins (test set 2 of Feig et al.[28]). Test set 2 contains nonhomologous, single-chain proteins that range from small protein fragments to large ones with over 800 residues and cover diverse native folds. The total electrostatic solvation free energies for all 22 proteins in test set 1 (which is the training set for numerical parametrization) are provided in Table 1 of the Supporting Information. Nina's radii were used in these calculations. The average error improves to less than 2% for GBSW/MS2 compared to that of over 6% for GBSW/MS. The maximum relative error of 9.6% was observed for a small 46-residue protein (PDB 1cbn). It is not clear why GBSW/MS and GBSW/MS2 appear to have particular difficulty for this protein. The maximum relative error is about 4% for the rest of the training set. In Figure 2, we compare the total solvation free energies calculated from GBSW/MS2 and PBEQ for test set 2. As will be shown later in this section, the CHARMM param22 vdW radii are nearly optimal for GBMV2. To make direct comparison to GBMV2 and also to test the sensitivity of the above numerical parametrization to (small) changes in input radii, the CHARMM param22 vdW radii were used in the calculations shown in Figure 2. The apparent correlation between PB and GB results is excellent, with a correlation coefficient of $R = 0.997$. The maximum and average absolute and relative errors are 708.2 and 47.7 kcal/mol, compared to those of 482.8 and 28.1 kcal/mol of GBMV2 (with default options and SHIFT −0.102 SLOPE 0.9085 P6 8). Interestingly, the GBSW/MS2 solvation free energies are actually better correlated with the GBMV2 results than with PBEQ/MS,

Table 1. Optimized Atomic Input Radii for GBSW/MS2 and GBMV2 in Comparison with the Original CHARMM param22 vdW and Nina's Radii[a]

group	atom[b]	vdW	Nina	GBSW/MS2	GBMV2[c]
hydrogen		varies	0.0	0.0	–
backbone	C	2.06	2.04	2.06	2.06
	O	1.70	1.52	1.75	1.70
	N	1.85	2.23	1.95	1.85
	CA	2.06	2.86	2.86	–
side chains					
all[d]	CB	2.175	2.67	2.80	–
all[d]	CD, CG	2.275	2.46	2.50	–
Lys	NZ	1.85	2.13	2.13	1.85
	CE	2.175	2.80	2.60	2.80
Arg	NH*	1.85	2.13	2.05	1.80
	NE	1.85	2.13	2.10	1.80
	CZ	2.00	2.80	2.60	2.80
Glu[e]	OE*	1.70	1.42	1.78	1.70
Gln[f]	NE2	1.85	2.15	2.10	1.95
	OE1	1.70	1.42	1.85	1.70
His[g]	ND1, NE2	1.85	2.31	2.10	1.85
	CD2, CE1	1.80	1.85	2.20	–
Trp	NE1	1.85	2.40	2.05	1.85
	C* (ring)	1.8–2.0	1.78	2.30	–
Tyr	OH	1.77	1.85	1.73	1.77
Ser, Thr	OG*	1.77	1.64	1.70	1.77
Pro	CB, CG, CD	2.175	1.98	2.20	–
Phe,Tyr	C* (ring)	1.99	2.00	2.30	–
methyl carbons[h]		2.06	2.44	2.70	–

[a] The CHARMM param22 vdW radii are used for all atoms not shown in this table. All values are in angstroms. [b] * is a Wild card character. [c] Entries with "–" denote cases where the radii have not been optimized for GBMV2 and the default CHARMM para22 vdW radii are used. [d] Unless otherwise specified. [e] Also for OD* of Asp and carbonyl oxygen atoms in the charged C-terminus. [f] Also for ND2 and OD1 of Asn. [g] Also for protonated His (Hsp). [h] Except CB of Ala.

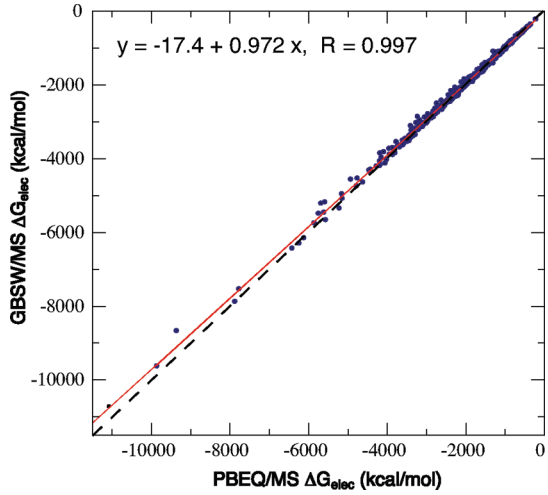

Figure 2. Comparison of the total electrostatic solvation free energies computed using GBSW/MS2 and PBEQ for all 611 proteins in test set 2. The dashed line plots the diagonal, and the red line corresponds to the best linear fit. The CHARMM para22 vdW radii were used as the input radii in both GB and PB calculations.

with $R = 0.999$. We note that, while the average error of GBMV2 is similar to what was previously reported,[28] the maximum GBMV2 error is higher in the current calculations. This might be attributed to slight difference in the PBEQ setup (Feig et al.[28] used a slightly larger grid of 0.25 Å). In addition, several of our PBEQ calculations of the total self-solvation free energies appeared to be unstable, where multiple calculations with different protein orientations failed to yield similar results and some of them even completely failed in reaching convergence within the maximum number of iterations. Changing grid specifications and other PBEQ parameters did not appear to resolve such occasional instabilities in a consistent fashion. Nonetheless, the average error and correlation coefficient reported above are minimally impacted by the presence of several less reliable PB results.

3.2. Initial Optimization of Input Radii: PMFs and Solvation Free Energies. Initial calculations using various sets of input radii suggested that neither Nina's radii[74,75] nor a modified set optimized for GBSW[13] offers an obvious advantage as the starting point for optimizing the GBSW/MS2 protein force field. Instead, initial optimization of the GBMV2 protein force field suggested that the CHARMM param22 vdW radii are nearly optimal for GBMV2. Therefore, the CHARMM param22 vdW radii were specified as the default for GBSW/MS2, and input radii for selected atoms were systematically optimized to reproduce the stabilities of pairwise and multibody polar and nonpolar interactions among backbone and side chain analogues in various poses in TIP3P. Even with nearly 50 PMFs, the parametrization appears to be underdetermined, and many radius sets similarly reproduce the TIP3P results in terms of the root-mean-squared deviation (rmsd) of the stabilities. The input radius set with the fewest modifications from the initial values was chosen to avoid overparametrization. The input radii of key backbone atoms were further co-optimized with the peptide backbone torsion energetics on the basis of peptide simulations (see the next section). We have also optimized the input radii for peptide backbone and polar side chains for GBMV2 to establish a fair benchmark for assessing GBSW/MS2. The final optimized input radii are summarized in Table 1. The associated CHARMM input files for setting up the atomic input radii for GBSW/MS2 and GBMV2 are provided in the Supporting Information. With these radii, GBSW/MS2 is able to reproduce the stabilities of polar pairwise interactions in TIP3P to about 1.1 kcal/mol rmsd (excluding two problematic pairs noted below) and those of nonpolar interactions to about 0.42 kcal/mol rmsd. The strengths of all interactions studied in this work in TIP3P, GBSW, GBMV2, and GBSW/MS2 are summarized Figures S1 and S2 of the Supporting Information.

Figure 3 compares the PMFs of nine representative pairwise interactions between various polar moieties using sets of optimized radii specific to each of the three implicit solvent models compared. These PMFs clearly demonstrate the ability of GBSW/MS2 to capture both the location and magnitude of the first desolvation peaks in the TIP3P PMFs to a degree that is comparable to that of GBMV2. This is in contrast to the original GBSW model, where the desolvation peaks are largely absent for all the pairs examined. A correct description of the desolvation peaks strongly supports that the proposed empirical parametrization is indeed capable of capturing a realistic MS-like solute–solvent boundary at the atomic level, despite the use of atomic-centric dielectric functions. Through input radius optimization, GBMV2 and

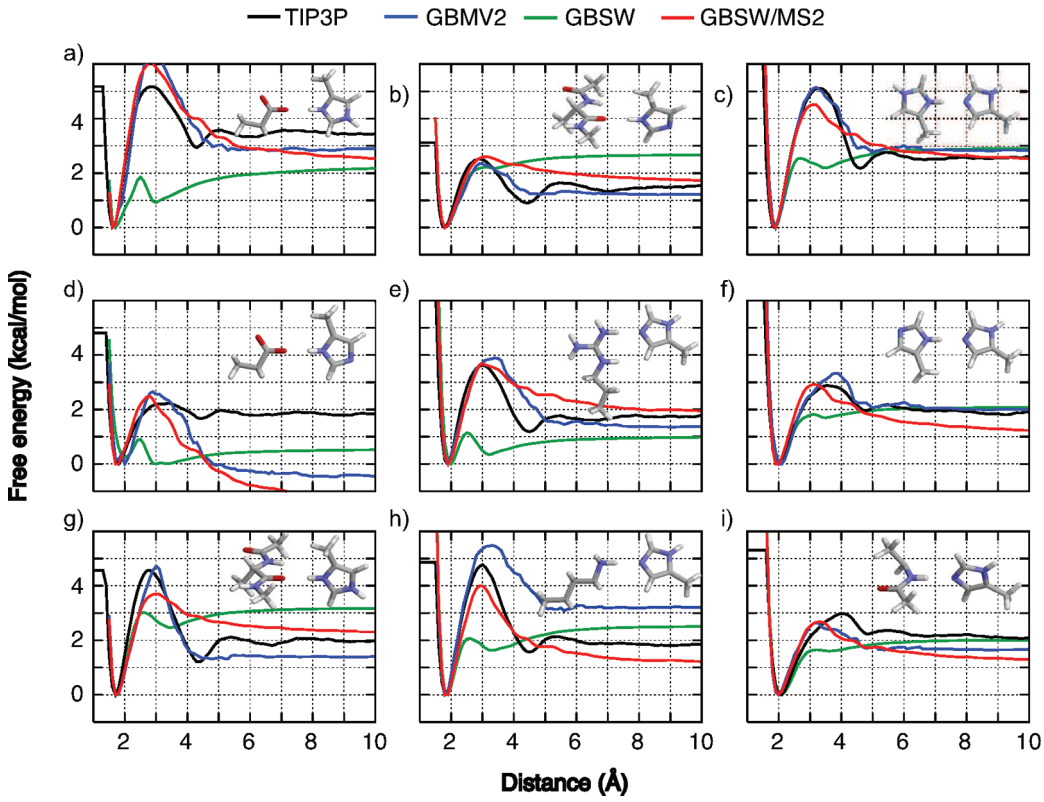

Figure 3. Comparison of the PMFs of nine representative pairwise interactions between polar moieties: (a) Hsp−Glu (hpe), (b) His−backbone carbonyl (hbco), (c) Hsp−His (hhp), (d) His−Glu (he), (e) His−Arg (hr), (f) His−His (hh), (g) Hsp−backbone carbonyl (hpbco), (h) His−Lys (hk), and (i) His−backbone amide (hb). All PMFs are aligned at the contact minimum. Details of the TIP3P PMF calculations were described in a previous paper.[13] PMFs in GBSW were computed using a previously optimized input radius set,[13] while those in GBSW/MS2 and GBMV2 were calculated using the optimized radii listed in Table 1.

GBSW/MS2 are able to reproduce the TIP3P stabilities of most pairwise interactions well. However, both GBMV2 and GBSW/MS2 appear to have difficulty in modeling Glu (and Asp), particularly its interaction with His (Figure 3d) or Lys (see Figure S1 in the Supporting Information). There is an unusual sensitivity to the choice of input radii, and no set was able to reproduce all related PMFs. The specific reason is not fully understood at this point. Interestingly, the final optimized radii yield accurate solvation free energies for side chain analogues of Asp/Glu (kcal/mol): −79.95/−78.03 (GBMV2) and −81.02/−78.76 (GBSW/MS2) compared to −80.62/−80.54 (TIP3P[76]) and −80.65/−79.12 (experimental[76]). The rms deviation from the TIP3P stabilities for all pairs except ek and he (see Figure S1) is about 1 kcal/mol for both GBSW/MS2 and GBMV2. Closer inspection reveals that GBMV2 PMFs tend to be rugged, which will lead to unstable forces and, consequently, numerical instability and poor energy conservation as previously observed in GBMV2 protein simulations.[58] The PMFs in GBSW/MS2 also contain small fluctuations. However, these fluctuations are smooth, and the atomic forces from GBSW/MS2 are as stable as in GBSW. Furthermore, as shown in Figure 4, the ability of GBSW/MS2 to reproduce the experimental solvation free energies of amino acid side chains is not compromised even though the input radius optimization was based on PMFs alone.

Previous analyses have strongly suggested that SA-based models have important limitations in describing the protein conformational dependence of nonpolar solvation.[23,24] In particular, it has been argued that both the solvent screening of the solute−solute dispersion interactions and length-scale dependence of hydrophobic solvation need to be properly described.[25] Ongoing development of more sophisticated nonpolar solvation models is beyond the scope of the current work. Nonetheless, it is important to examine the ability of the simple SA model in reproducing the TIP3P PMFs of nonpolar interactions. The results of all 25 nonpolar interactions examined in this work are summarized in Figure S2 of the Supporting Information. While there remains a systematic underestimation of the strengths of three-body hydrophobic associations (e.g., fff and lll interactions in Figure S2) general to simple SA models,[24] GBSW/MS2 with SA can be parametrized to reproduce the TIP3P results to an overall rmsd of 0.42 kcal/mol. Figure 5 plots four representative nonpolar PMFs calculated in TIP3P, GBSW, GBMV, and GBSW/MS2 solvents. All three implicit solvents are able to more or less capture the overall stabilities of these interactions. However, regardless of adopting vdW or MS-like surfaces, none of these (SA) models are able to reproduce the desolvation peaks observed in TIP3P. This further illustrates the insufficiency of SA-based models to capture the delicate balance between cavitation (creating a hydrophobic cavity within the solvent) and solute−solvent dispersion interactions and, particularly, the conformational dependence of this balance.

3.3. Conformational Equilibria of Model Peptides. Examination of interactions between backbone and side chain moieties described above clearly demonstrates the challenge in

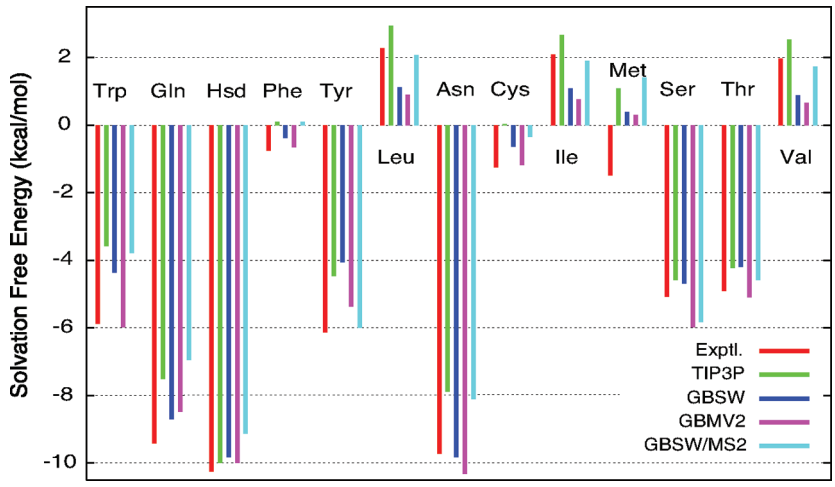

Figure 4. Experimental and calculated total solvation energies of amino acid side chain analogues. The experimental and TIP3P values were taken from ref 63. The GBSW results were computed using a previously optimized set,[13] and the GBMV2 and GBSW/MS2 results were computed using the radii listed in Table 1. All model compounds have the default geometries as defined in the CHARMM param22 fore field.[42] The rms deviations from the experimental results are 1.41, 1.11, 1.07, and 1.34 kcal/mol for TIP3P, GBSW, GBMV2, and GBSW/MS2, respectively.

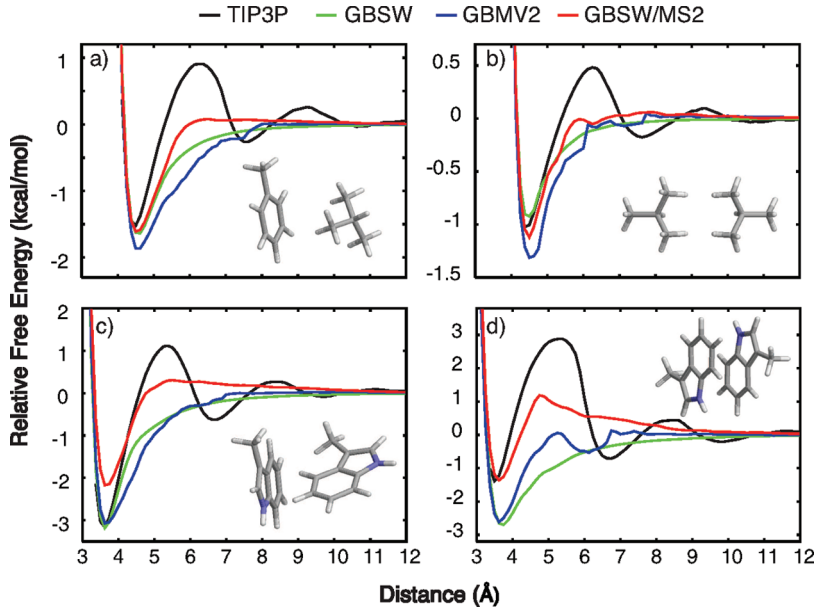

Figure 5. Comparison of the PMFs of four representative pairwise interactions between nonpolar moieties: (a) Phe−Leu (fl), (b) Leu−Leu (head-to-head, ll_h), (c) Trp−Trp (edge-to-face, ww_etf), and (d) Trp−Trp (antiparallel displaced, ww_apd). All PMFs are aligned at the largest separation distances. Details of the free energy protocol for calculating these TIP3P PMFs were described in previous papers.[24,25] PMFs in GBSW were computed using a previously optimized input radius set,[13] while those in GBSW/MS2 and GBMV2 were calculated using the radii listed in Table 1.

accurately balancing competing interactions even at the small-molecule level. It is thus important to achieve sufficient cancellation of errors at peptide and protein levels, such as via iterative optimization of important protein parameters, particularly backbone torsional energetics, together with the solvation model. A large number of folding and control REMD simulations (>100) of the model peptides were carried out to explore various parameter sets. The final optimized GBSW/MS2 protein force field achieves reasonable success in balancing the secondary structural propensities and in recapitulating the experimental structure and stabilities of the five model peptides used in this work, but nonetheless with important limitations (detailed later in the discussion). Figure 6 examines the average helicity of $(AAQAA)_3$ as a function of temperature for several representative combinations of backbone input radii and CMAP adjustments. The calculated helicity appears to converge well with 20 ns REMD simulations, as illustrated by comparing results from two independent simulations for one of the parameter sets explored (thick and thin solid red traces in Figure 6). Clearly, approximation of an MS-like dielectric boundary significantly increases the intrinsic propensity to form an ideal α-helix. For example, GBSW/MS2 yields an extremely stable helix with a melting temperature of ~500 K (solid green trace in Figure 6), even though the corresponding backbone hydrogen-bonding strength estimated from a modified alanine dipeptide dimer model is only 1.55 kcal/mol, weaker than that in either TIP3P

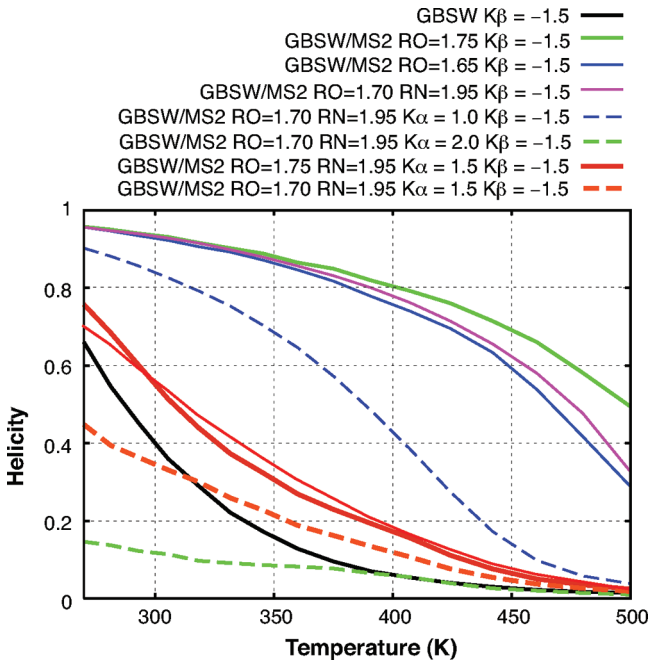

Figure 6. Total helicity of $(AAQAA)_3$ as a function of temperature in GBSW/MS2 implicit solvent with different combinations of peptide backbone torsion modifications and input radii. The result calculated using a previously optimized GBSW protein force field[13] is also shown for comparison. The helicity was calculated from the averaged frequency of 1−4 hydrogen bonding, defined by $d_{O_i}HN_{i+4} \leq 2.6$ Å, during the second halves of 20 ns REMD simulations included in computing the averages. The backbone torsion modifications were implemented using eq 10 with various k^α and k^β values. RO and RN denote the input radii for backbone carbonyl oxygen atoms and amide nitrogen, respectively. RN = 2.05 Å when not explicitly specified. The stabilities of the backbone hydrogen bonding for a modified alanine dipeptide dimer[12] are (kcal/mol) 1.9 ± 0.3 (TIP3P), 1.80 (GBSW), 1.55 (GBSW/MS2, RO = 1.75 Å), 0.94 (GBSW/MS2, RO = 1.65 Å), 1.44 (GBSW/MS2, RO = 1.75 Å, RN = 1.95 Å), and 1.1 (GBSW/MS2, RO = 1.70 Å, RN = 1.95 Å). Results from two independent REMD simulations are shown for GBSW/MS2 with the final selected parameters to illustrate convergence (thick and thin solid red traces).

(1.9 ± 0.3 kcal/mol) or the previously optimized GBSW solvent (1.8 kcal/mol).[13] Similar observations can be made with GBMV2, as illustrated in Figure S3 of the Supporting Information. This dramatic difference between two types of dielectric boundaries might be rationalized by a significant reduction of internal high dielectric regions (and thus reduced dielectric screening) with an MS compared to a vdW-like surface. Importantly, the intrinsic helical propensity with the MS is strong and cannot be overcome by tuning the backbone hydrogen-bonding strength alone. For example, the stability of $(AAQAA)_3$ remains grossly overestimated even if one reduces the backbone hydrogen-bonding strength (in the model dimer) to less than 1 kcal/mol (solid blue trace in Figure 6). Instead, it is necessary to modify the peptide backbone torsion energetics, such as by directly reducing the stability of helical conformations. It turns out that k^α up to 2 kcal/mol is sufficient to completely destabilize the helical conformation. The optimal choice appears to be $k^\alpha = 1.5$ kcal/mol and $k^\beta = -1.5$ kcal/mol (this choice of k^β was based on further simulations of β-hairpins). Given this backbone torsion modification, input radii of key backbone atom types (carbonyl oxygens and amide nytrogens) can be slightly adjusted to fine-tune the balance between helical and random coil/extended conformations, as illustrated by the solid and dashed red traces in Figure 6.

Disappointedly, further fine-tuning of the backbone input radii and ϕ/ψ torsion energetics based on conformational equilibria of the GB1p series hairpins has been met with much greater difficulty compared to our previous efforts in optimizing the GBSW protein force field.[13] The primary challenge is an apparent inability to achieve convergence on conformational equilibria of GB1p-derived hairpins, even with REMD simulations up to 150 ns. Figure 7 compares the energetic and structural properties of the lowest temperature ensembles (at 270 K) from 50 ns control and folding simulations of GB1m3, the most stable hairpin in the series. While several replicas did successfully fold during the folding simulation (panel c), control and folding runs do not sample similar regions of the conformational space, as reflected in the probability distributions of the number of native hydrogen bonds (panel d). There appear to be significant energetic barriers separating the fully folded basin from nearly folded and unfolded regions, such that the low-energy basin predominantly sampled in the control simulation is not visited during the folding run (panels a and b). One can speculate that, to arrive at (or escape) the fully folded native basin, multiple (polar) interactions need to be rearranged at the same time, which likely involves significant collective (desolvation) barriers due to the MS-like underlying dielectric boundary. In addition, conformational diffusion is expected to be slower in GBSW/MS2 (or GBMV2) than in GBSW due to the presence of a desolvation barrier, which also makes sampling more difficult in general.

Similar difficulty was experienced in achieving convergence in conformational equilibria for other GB1p hairpins. The control REMD simulations systematically overestimate the stabilities, while the folding runs systematically underestimate the stabilities. This is clearly reflected by comparing the probability distributions of the number of native hydrogen bonds in Figure 8. HP5A is the single case where a certain level of apparent convergence was achieved. Similar to previous simulations with GBSW,[13] GB1p is the most challenging sequence to simulate with a more flexible turn (DATK vs PATG in HP5A and GB1m3). No replica successfully folded in any of the multiple 50 ns REMD folding simulations (e.g., see "folding-1" in Figure 8a). The only folding event was observed during a 150 ns REMD simulation with a smaller temperature range of 270−400 K ("folding-2" in Figure 8; the folding event occurred around the 50 ns mark during the simulation). For GB1m1, no replica ever folded during multiple 50 ns REMD folding simulations with various temperature ranges. However, several replicas remained folded even when the simulation was extended to 100 ns (with a temperature range of 270−500 K) (see Figure 8d). Again, the difficulty in achieving convergence is clearly related to a significantly more rugged energy landscape with an MS-like underlying dielectric boundary. It also reflects important limitations of standard REMD in enhancing conformational sampling and re-emphasizes the importance of careful

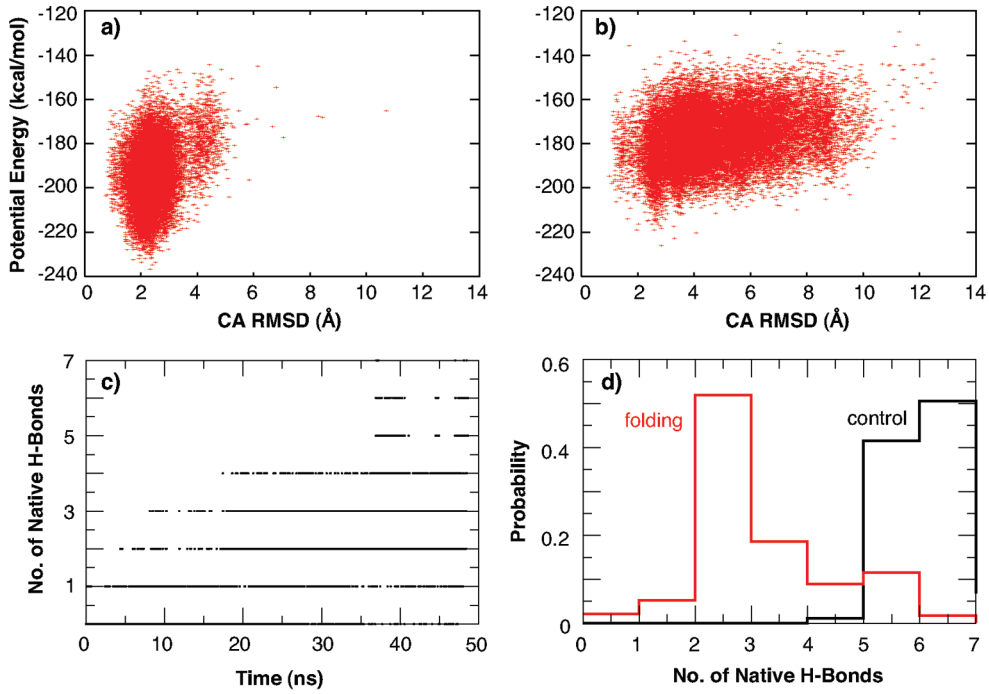

Figure 7. REMD simulations of the gb1m3 hairpin. Potential energy and C_α rmsd scatter plots for all conformations sampled at 270 K from (a) control and (b) folding REMD simulations, (c) the number of native hydrogen bonds as a function of simulation time, and (d) probability distributions of the number of native hydrogen bonds at 270 K. The probability distributions were computed from the last 20% (10 ns) of 50 ns REMD control and folding simulations. GBSW/MS2 was used with the atomic input radii as specified in Table 1, together with a modified CMAP (eq 10 with $k^\alpha = 1.5$ kcal/mol and $k^\beta = -1.5$ kcal/mol). The native hydrogen bonds include (in protein GB1 residue numbering) E42(N)−T55(O), E42(O)−T55(N), T44(N)−T53(O), T44(O)−T53(N), D46(N)−T51(O), D46(O)−T51(N), and D47(O)−K50(N).

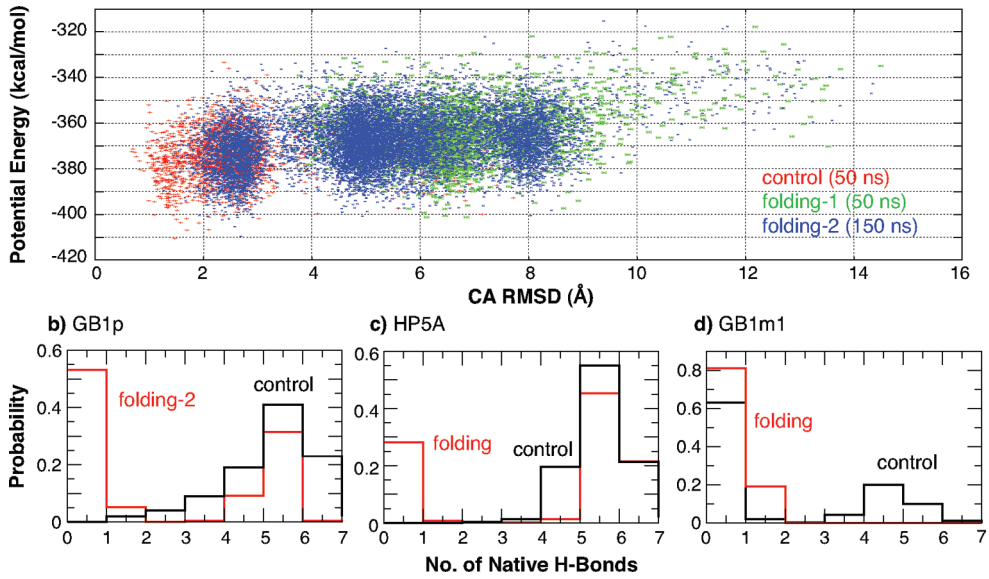

Figure 8. REMD simulations of GB1p, HP5A, and GB1mp: (a) potential energy and C_α rmsd scatter plots for all conformations sampled at 270 K from multiple control and folding simulations of GB1p; (b−d) probability distributions of the number of native hydrogen bonds at 270 K. Details of the simulation and analysis are provided in the caption of Figure 7, except as otherwise noted here. Two folding REMD simulations are shown for GB1p, where the temperature range is 270−500 K for "folding-1" and 270−400 K for "folding-2". The folding simulation of HP5A was also carried out with a temperature range of 270−400 K. The length of the GB1m1 control simulation was 100 ns.

examination of convergence by independent runs initiated from distal points in the conformational space. Nevertheless, results from these folding and control simulations suggest that the optimized GBSW/MS2 force field provides a reasonable balance between helical and coil/extended conformations, being able to fold both helices and β-hairpins. More importantly, albeit inconclusive due to a lack of satisfactory convergence, results shown in Figures 7 and 8 indicate that GBSW/MS2 roughly captures the trend in stabilities of GB1p series hairpins, GB1m3 > GB1p ≈ HP5A > GB1m1.

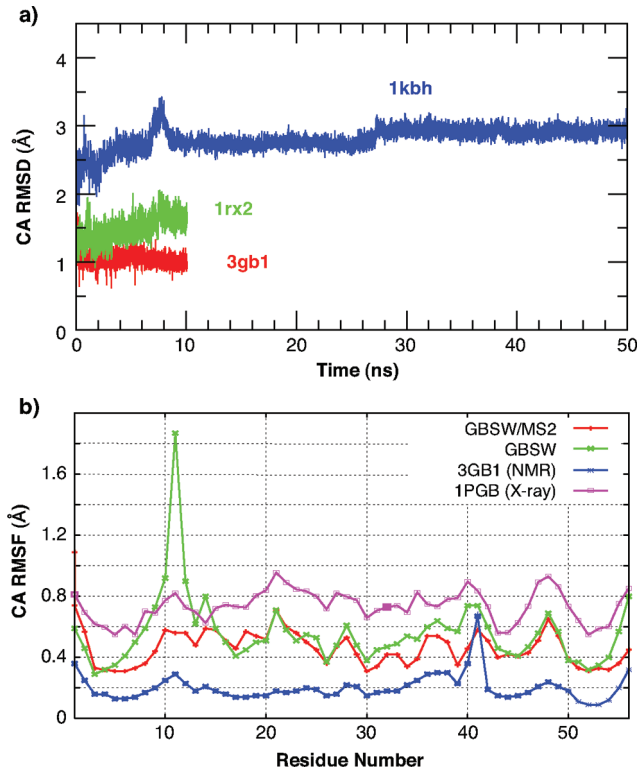

Figure 9. (a) C_α rmsd from the PDB structures during 10−50 ns simulations of two proteins (1rx2 and 3gb1) and one protein complex (1kbh) at 300 K and (b) C_α RMSF calculated from the second halves of 10 ns MD simulations of 3gb1 in GBSW/MS2 and GBSW implicit solvents in comparison with values calculated from the NMR and X-ray PDB structures. The disorder termini in complex 1kbh were not included in the rmsd calculation. The C_α RMSF of PDB 3gb1 was calculated using all 32 models of the NMR ensemble, and that of the X-ray structure (PDB 1pgb) was converted from the *B* factors using the relationship *B* factor $= 8\pi^2(\text{RMSF})^2/3$.

3.4. Control Simulation of Peptides and Proteins. Control simulations of two proteins and a protein complex were also carried out to confirm the suitability of simulating the native states using the optimized GBSW/MS2 force field, particularly considering that the previous parametrization (GBSW/MS) appears to have a problem stabilizing several proteins. As shown in Figure 9, all these systems remain stable throughout the simulation time scales up to 50 ns for the CBP/ACTR complex (PDB 1kbh). Note that both CBP and ACTR domains are intrinsically disordered in unbound states and the complex is believed to retain substantial flexibility.[77,78] In Figure 9b, we compare the root-mean-square fluctuations (RMSFs) computed from 10 ns simulations in GBSW/MS2 and GBSW together with those obtained from NMR and X-ray PDB structures. Even though the first hairpin turn (near residue 11) appears to be substantially more flexible in GBSW than in GBSW/MS2, there is little experimental evidence for this from either the *B* factors or the NMR ensemble. Interestingly, the RMSF profile from the GBSW/MS2 simulation better tracks the one converted from the *B* factors, with a correlation of 0.66 compared to that of 0.41 for GBSW. These control simulations suggest that GBSW/MS2 provides a suitable alternative to GBMV2 (or explicit solvent) for protein simulations with enhanced stability and computational efficiency.

The ability of the GBSW/MS2 force field to describe unfolded protein states is examined by comparing the calculated NMR scalar coupling constants with experimental values for Ala_5. The results are summarized in the Supporting Information, Figure S4. Deviations from the experimental values are comparable to previous analysis of other force fields.[73,79,80] Using the same criteria for classifying the backbone (ϕ, ψ) space as those used by Best et al.,[80] the populations of α, β, and PPII regions are 7.7%, 69%, and 22% at 300 K, respectively. Interestingly, the calculated helical population using the optimized GBSW/MS2 force field is one of the smallest among all force fields reported (e.g., compared to 41.5% from CHARMM27/CMAP with TIP3P water) and is right within the range of helicities estimated by reweighting various basins on the basis of NMR scalar coupling constants.[80] However, the β population appears to be significantly overestimated, at the expense of lower PPII conformation. This is likely a consequence of the optimization strategy that directly targets the balance between α/β secondary structure folding propensities. Additional attention to the details of unfolded states will need to be addressed in future studies.

4. Conclusions

MS is widely considered as the most proper choice for representing the solute−solvent boundary in continuum electrostatics-based implicit solvent models such as GB and PB, yet most GB models are based on atom-centered dielectric functions that describe vdW-like surfaces for computational efficiency and numerical stability. A surprisingly simple, yet effective, approximation to true MS is described here in the context of GBSW implicit solvent. By directly optimizing appropriate *numerical parameters*, the sum of a series of CFA and higher order terms was shown to be capable of nicely approximating the true MS. The new model, termed GBSW/MS2, is able to reproduce the exact MS PB-derived effective Born radii with a correlation of 0.95, apparently second only to the GBMV2 model among the few existing MS-based GB models. More importantly, examination of the PMFs of pairwise polar interactions demonstrates that desolvation barriers, a key signature of a true MS, are properly captured in GBSW/MS2. GBSW/MS2 fully retains the computational efficiency and numerical stability of the original GBSW model, allowing one to extensively optimize the *physical parameters* to obtain a balanced GBSW/MS2 protein force field. The optimization was guided by PMFs of over 40 polar and nonpolar interactions and relied extensively on direct simulation of the conformational equilibria of a set of carefully chosen helical and β-hairpin peptides. The key physical parameters optimized include the atomic input radii for specifying the location of solute−solvent boundary and peptide backbone torsion energetics. The final optimized GBSW/MS2 protein force field not only satisfactorily captures the delicate balance between solvation and intramolecular interactions on the model compound level compared to the TIP3P solvent, but also appears to reasonably describe the balance between helical and coil/β conformations on the model level.

Nonetheless, several important limitations have prevented one from further fine-tuning the GBSW/MS2 force field. In particular, the presence of desolvation peaks due to MS significantly increases the ruggedness of the underlying potential energy surface and hinders expedient conformational sampling. Satisfactory convergence in simulated conformational ensembles could not be achieved for the β-hairpins using REMD simulations up to 150 ns with various temperature ranges. In contrast, good convergence was achieved for the same set of hairpins within 30−50 ns for GBSW with a vdW-like underlying dielectric boundary.[13] This also highlights the importance of improved understanding of the efficacy of REMD[71] and developing other enhanced sampling techniques[81,82] in development and optimization of (implicit solvent) protein force fields. The current work focuses on the electrostatic component of the solvation mode,l and the nonpolar solvation free energy is still estimated using the simple SA model. While reasonable success can be achieved in reproducing stabilities of pairwise and multibody nonpolar interactions in TIP3P, it must be emphasized that SA-based models have important limitations in capturing the conformational dependence of solvation.[23,25] This deficiency is partially reflected in the lack of fine features in PMFs of nonpolar interactions, using either MS or vdW-like dielectric boundaries. Development of more sophisticated nonpolar solvation models will be necessary to achieve a fully balanced implicit-solvent-based force field for reliable simulation of peptide and protein conformational equilibria. This is a formidable task and will also depend critically on further improvement in the underlying protein models.[1,2,83] The GBSW/MS2 approximation introduced here not only provides a viable alternative for modeling protein structure and interactions, but is also expected to constitute an important step toward the development of an efficient, stable, and fully balanced GB-based implicit solvent protein force field.

Acknowledgment. I am in debt to Dr. Wonpil Im for providing the protein test sets used in this work, and I thank Drs. Wonpil Im and Debabani Ganguly for many helpful discussions. Mr. Weihong Zhang carried out the simulations of CBP/ACTR in various GBSW and GBSW/MS2 force fields. This work was supported by an Innovative Research Award from the Terry C. Johnson Center for Basic Cancer Research and a CAREER Award from NSF (Grant MCB 0952514). This paper is Contribution No. 11-007-J from the Kansas Agricultural Experiment Station.

Supporting Information Available: Total electrostatic solvation free energies of 22 proteins in test set 1, stabilities of all 44 polar and nonpolar interactions, comparison of helicities of $(AAQAA)_3$ in GBSW and GBMV2, summary of $(Ala)_5$ backbone conformational equilibria, and CHARMM input stream files for setting up optimized GBSW/MS2 and GBMV2 implicit solvent. This information is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Ponder, J. W.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(2) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584–1604.

(3) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.

(4) Feig, M.; Brooks, C. L., III. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.

(5) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(6) Chen, J.; Brooks, C. L., III; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.

(7) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(8) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(9) Gallicchio, E.; Paris, K.; Levy, R. M. *J. Chem. Theory Comput.* **2009**, *5*, 2544–2564.

(10) Lee, M. S.; Olson, M. A. *J. Phys. Chem. B* **2005**, *109*, 5223–5236.

(11) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.

(12) Im, W.; Chen, J.; Brooks, C. L., III. *Adv. Protein Chem.* **2005**, *72*, 173–198.

(13) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.

(14) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.

(15) Jang, S.; Kim, E.; Pak, Y. *Proteins* **2007**, *66*, 53–60.

(16) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins* **2002**, *46*, 24–33.

(17) Lazaridis, T. *Proteins* **2005**, *58*, 518–527.

(18) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.

(19) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435–445.

(20) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.

(21) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

(22) Constanciel, R.; Contreras, R. *Theor. Chim. Acta* **1984**, *65*, 1–11.

(23) Levy, R. M.; Zhang, L. Y.; Gallicchio, E.; Felts, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 9523–9530.

(24) Chen, J.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 2444–2445.

(25) Chen, J.; Brooks, C. L., III. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.

(26) Born, M. *Z. Phys.* **1920**, *1*, 45–48.

(27) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.

(28) Feig, M.; Onufriev, A.; Lee, M.; Im, W.; Case, D.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 265–284.

(29) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(30) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.

(31) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005–3014.

(32) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098–8106.

(33) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983–10990.

(34) Dominy, B. N.; Brooks, C. L., III. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.

(35) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theor. Chem. Acc.* **1999**, *101*, 426–434.

(36) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(37) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606–10614.

(38) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691–1702.

(39) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.

(40) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008–3022.

(41) Haberthur, U.; Caflisch, A. *J. Comput. Chem.* **2008**, *29*, 701–715.

(42) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(43) Feig, M.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Phys. Chem.* **2003**, *107*, 2831–2836.

(44) MacKerell, A., Jr.; Feig, M.; Brooks, C., III. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(45) MacKerell, A., Jr.; Feig, M.; Brooks, C., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(46) Khandogin, J.; Chen, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546–18550.

(47) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.

(48) Khandogin, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880–16885.

(49) Ganguly, D.; Chen, J. *J. Am. Chem. Soc.* **2009**, *131*, 5214–5223.

(50) Chen, J. *J. Am. Chem. Soc.* **2009**, *131*, 2088–2089.

(51) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.

(52) Lu, Q.; Luo, R. *J. Chem. Phys.* **2003**, *119*, 11035.

(53) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769–14772.

(54) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.

(55) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.

(56) Yu, Z. Y.; Jacobson, M. P.; Friesner, R. A. *J. Comput. Chem.* **2006**, *27*, 72–89.

(57) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.

(58) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719–729.

(59) Tjong, H.; Zhou, H. X. *J. Phys. Chem. B* **2007**, *111*, 3055–3061.

(60) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(61) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

(62) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151–176.

(63) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740–5761.

(64) Shalongo, W.; Dugad, L.; Stellwagen, E. *J. Am. Chem. Soc.* **1994**, *116*, 8288–8293.

(65) Blanco, F. J; Rivas, G.; Serrano, L. *Nat. Struct. Biol.* **1994**, *1*, 584–590.

(66) Fesinmeyer, R. M.; Hudson, F. M.; Andersen, N. H. *J. Am. Chem. Soc.* **2004**, *126*, 7238–7243.

(67) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.

(68) Lei, H.; Duan, Y. *Curr. Opin. Struct. Biol.* **2007**, *17*, 187–191.

(69) Nymeyer, H. *J. Chem. Theory Comput.* **2008**, *4*, 626–636.

(70) Denschlag, R.; Lingenheil, M.; Tavan, P. *Chem. Phys. Lett.* **2008**, *458*, 244–248.

(71) Zheng, W.; Andrec, M.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 15340–15345.

(72) Periole, X.; Mark, A. E. *J. Chem. Phys.* **2007**, *126*, 014903.

(73) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.

(74) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

(75) Nina, M.; Im, W.; Roux, B. *Biophys. Chem.* **1999**, *78*, 89–96.

(76) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.

(77) Demarest, S. J.; Martinez-Yamout, M.; Chung, J.; Chen, H. W.; Xu, W.; Dyson, H. J.; Evans, R. M.; Wright, P. E. *Nature* **2002**, *415*, 549–553.

(78) Ebert, M. O.; Bae, S. H.; Dyson, H. J.; Wright, P. E. *Biochemistry* **2008**, *47*, 1299–1308.

(79) Wickstrom, L.; Okur, A.; Simmerling, C. *Biophys. J.* **2009**, *97*, 853–856.

(80) Best, R. B.; Buchete, N.; Hummer, G. *Biophys. J.* **2008**, *95*, 4494.

(81) Gao, Y. Q. *J. Chem. Phys.* **2008**, *128*, 064105.

(82) Liwo, A.; Czaplewski, C.; Oldziej, S.; Scheraga, H. A. *Curr. Opin. Struct. Biol.* **2008**, *18*, 134–139.

(83) Kang, M.; Smith, P. E. *J. Comput. Chem.* **2006**, *27*, 1477–1485.

CT100251Y