JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Continuous Constant pH Molecular Dynamics in Explicit Solvent with pH-Based Replica Exchange

Jason A. Wallace and Jana K. Shen*

Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019, United States

ABSTRACT: A computational tool that offers accurate pK_a values and atomically detailed knowledge of protonation-coupled conformational dynamics is valuable for elucidating mechanisms of energy transduction processes in biology, such as enzyme catalysis and electron transfer as well as proton and drug transport. Toward this goal we present a new technique of embedding continuous constant pH molecular dynamics within an explicit-solvent representation. In this technique we make use of the efficiency of the generalized-Born (GB) implicit-solvent model for estimating the free energy of protein solvation while propagating conformational dynamics using the more accurate explicit-solvent model. Also, we employ a pH-based replica exchange scheme to significantly enhance both protonation and conformational state sampling. Benchmark data of five proteins including HP36, NTL9, BBL, HEWL, and SNase yield an average absolute deviation of 0.53 and a root mean squared deviation of 0.74 from experimental data. This level of accuracy is obtained with 1 ns simulations per replica. Detailed analysis reveals that explicit-solvent sampling provides increased accuracy relative to the previous GB-based method by preserving the native structure, providing a more realistic description of conformational flexibility of the hydrophobic cluster, and correctly modeling solvent mediated ion-pair interactions. Thus, we anticipate that the new technique will emerge as a practical tool to capture ionization equilibria while enabling an intimate view of ionization coupled conformational dynamics that is difficult to delineate with experimental techniques alone.

INTRODUCTION

Solution pH has a profound effect on the stability and the function of proteins by changing the protonation states of titratable groups. Proteins can become denatured under extreme pH conditions. Enzymes are often catalytically active in a narrow pH range.[1] Protein−protein interactions[2] and protein−ligand binding[3] are also modulated by the protonation states of titratable groups. Accurate determination of active-site pK_a values informs about the catalytic mechanism of proteins.[4] Knowledge of the native- and denatured-state pK_a values can be used to quantify electrostatic effects on protein stability.[5]

Although the importance of solution pH has long been recognized, molecular simulation techniques have traditionally neglected it. In a standard molecular dynamics (MD) simulation the protonation states of ionizable side chains are set at the beginning of the simulation based on the comparison of the desired pH condition and the solution or model compound pK_a values. This fixed protonation scheme can be a source of error in several instances. For example, if the pK_a values are near the pH of interest, then the protonated and deprotonated states should coexist, which obviously is not reflected in simulation with fixed protonation states. Additionally, even when reasonable protonation states may be set for the initial conformation, conformational rearrangement may favor an entirely new set of protonation states.

In recent years, considerable effort has been made to develop methodologies that explicitly include pH as an external parameter in MD simulations, similar to temperature, allowing protonation states of ionizable groups to respond to changes in the chemical environment and the external pH.[6−11] These constant pH techniques differ in the way protonation states are updated. In the discrete methods, protonation states are periodically updated using Monte Carlo sampling, while in the continuous approach titration coordinates are introduced and propagated simultaneously with the spatial coordinates (see a most recent review).[12] One of the most promising constant pH techniques, termed continuous constant pH molecular dynamics (CPHMD)[9,13] is based on the λ dynamics approach to free energy calculations,[14] allowing ionizable groups to switch continuously between protonated and unprotonated forms. Protonation and deprotonation are accomplished in a manner similar to many free energy simulation techniques, where an alchemical coordinate, λ, is introduced. The novelty of the λ dynamics approach lies in the fact that the alchemical coordinate is assigned to a fictitious λ particle, and the force on the particles is derived analytically. CPHMD has been shown to give accurate and robust predictions for protein pK_a values[12] and has opened a door to theoretical studies of pH-dependent protein dynamics and folding.[15−17]

In the aforementioned CPHMD method, the generalized Born (GB) implicit-solvent model is used to calculate forces on both spatial and titration coordinates. The major advantage of using GB models in constant pH methodologies is that convergence of pK_a's can be achieved with a reasonable amount of sampling time, which has not been demonstrated feasible with explicit-solvent models (see more discussions later). Another benefit of using GB models within the CPHMD framework is that forces on the titration coordinates can be computed analytically. However, as CPHMD and other GB-based constant pH techniques are maturing into practical tools, problems inherited from the underlying GB models are becoming the limiting factor for further improvement of accuracy. Recent GB simulation studies have revealed several problems that seem

Received: February 28, 2011
Published: June 22, 2011

difficult to overcome. Specifically, attractive electrostatic interactions are overestimated,[18,19] and improvement through adjustment of GB input radii that define dielectric boundary[18,19] is limited.[20] Also, due to the lack of solvent granularity, GB simulations cannot reproduce the solvation peaks seen in the interaction free energy profiles from explicit-solvent simulations.[18] Furthermore, there have been noted problems with the stability of hydrophobic interactions[21,22] and the overly compact and rigid unfolded states,[23,24] which are likely due to the approximate nature of the nonpolar solvation term based on solvent-accessible surface area (SA model). Finally, the inaccuracies of the GB/SA model in the representation of electrostatic and nonpolar energetics result in a more favorable sampling of helical relative to extended states.[25,24]

The limitations of GB models affect the accuracy and the applicability of the CPHMD method in several ways. First, a small error in the electrostatic solvation energy calculated by the GB model alters the relative deprotonation free energy in reference to solution and therefore the pK_a shift. This type of "electrostatic" error is typically small for solvent-exposed residues, because the GB model, in particular GBSW used in this work, has been tuned to reproduce the explicit-solvent data of solvent-exposed polar or charged interactions.[18] However, the "electrostatic" error becomes significant for deeply buried residues[26,12] because the inaccuracy in the desolvation energies of deeply buried atoms remains an unsolved problem in GB models. Nonetheless, the electrostatic error is systematic (Wallace and Shen, unpublished data), and a post correction may be introduced if necessary. The second type of GB-related error which affects the accuracy of $\Delta\Delta G^{\mathrm{deprot}}$ arises from the small distortion in the conformation or distribution of conformations. The impact of this "conformational" error on the protonation-state sampling is typically not systematic, and the extent of the error is unpredictable. Finally, the dependence of conformational sampling on the GB model also hinders the application of the CPHMD method to polyionic systems, such as DNA and RNA, for which GB models are not well suited.

In light of the above considerations, we introduce here a method to extend the CPHMD framework to explicit-solvent simulations. In principle, forces on both spatial and titration coordinates can be derived from explicit-solvent sampling. However, the latter is not practical because a lengthy simulation time is required to accurately compute solvation-related forces based on explicit-solvent sampling. Consequently, we devise a method which takes advantage of the efficiency of the GB model to compute solvation forces on titration coordinates while propagating conformational dynamics via all-atom interactions in explicit solvent. Additionally, we implement a replica-exchange protocol based on the pH biasing energy to significantly accelerate the convergence of the simultaneous sampling of protonation and conformational space. Thus, by making use of the more accurate explicit-solvent sampling, the new method aims to improve the accuracy of CPHMD by reducing the aforementioned "conformational" error and to allow applications to many problems where implicit-solvent models are not feasible.

The rest of the paper is organized as follows. First, we describe in detail the CPHMD method in explicit solvent and the pH-based replica-exchange protocol. We then examine potential artifacts due to the use of both explicit- and implicit-solvent schemes and the response of solvent molecules to titration. Next, we present and discuss results of model compound titrations and analyze the convergence behavior with the new sampling protocol. Finally, we benchmark the accuracy of the new method by calculating pK_a values of five proteins including HP36, NTL9, BBL, HEWL, and SNase. We compare the results with the GB-based CPHMD simulations and experiment. We find that the explicit-solvent CPHMD offers slightly more accurate pK_a predictions but significantly deeper physical insights. Surprisingly, convergence of the explicit-solvent CPHMD titrations is achieved for all proteins with a simulation length of 1 ns per replica, suggesting that the new method will emerge as a powerful and practical tool for theoretical studies of electrostatic phenomena.

METHODS

Continuous Constant pH Molecular Dynamics in Implicit Solvent. To prepare for the discussion of the explicit-solvent based CPHMD method, we first briefly review the CPHMD method employing the generalized Born (GB) implicit-solvent model; more details can be found elsewhere.[9,13,26] Based on the λ dynamics technique,[14] CPHMD utilizes an extended Hamiltonian to simultaneously propagate spatial (real) and titration (virtual) coordinates. Thus, the total Hamiltonian of the system can be written as

$$\mathscr{H}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) = \sum_a \frac{m_a}{2}\dot{\mathbf{r}}_a^2 + U^{\mathrm{int}}(\{\mathbf{r}_a\}) + U^{\mathrm{hybr}}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) + \sum_i \frac{m_i}{2}\dot{\boldsymbol{\theta}}_i^2 + U^*(\{\boldsymbol{\theta}_i\}) \quad (1)$$

where $a = 1,N_{\mathrm{atom}}$ is the index for atomic coordinates, and $i = 1$, N_{titr} is the index for the continuous variables θ_i which is related to the titration coordinate λ_i by

$$\lambda_i = \sin^2(\theta_i) \quad (2)$$

Boundaries are naturally imposed on the titration coordinate through the sine function, where $\lambda_i = 0$ corresponds to the protonated state and $\lambda_i = 1$ corresponds to the deprotonated state. For residues with two competing titration sites, a second continuous variable can be included to allow interconversion between proton tautomeric states. This is indicated in eq 1 by bold $\boldsymbol{\theta}$.

In eq 1 the first term is the kinetic energy of the real system (atoms), U^{int} is the internal potential energy which is independent of titration, and U^{hybr} is a hybrid effective energy term which enables the coupling between conformational and protonation degrees of freedom. In the GB-based CPHMD method, it is written as a sum of the van der Waals, Coulombic, and GB electrostatic solvation free energies:

$$U^{\mathrm{hybr}}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) = U^{\mathrm{vdW}}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) + U^{\mathrm{Coul}}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) + U^{\mathrm{GB}}(\{\mathbf{r}_a\},\{\boldsymbol{\theta}_i\}) \quad (3)$$

where the latter is given as[27,28]

$$U^{\mathrm{GB}} = -\frac{1}{2}\sum_{a,b}\left(1-\frac{e^{-\kappa r_{ab}}}{\epsilon_w}\right)\frac{q_a q_b}{\sqrt{r_{ab}^2+\alpha_a\alpha_b\exp(-r_{ab}^2/F\alpha_a\alpha_b)}} \quad (4)$$

Here r_{ab} is the distance between two atoms, q_a and q_b are the respective partial charges, ϵ_w is the dielectric constant for water, and α_a and α_b are the effective Born radii, which can be interpreted as the spherically averaged distance from the atom to the dielectric boundary. To approximately account for the effect of salt screening, a Debye–Hückel term[29,28] is included in the above equation, where $\kappa^2 = 8\pi q^2 I/\epsilon k_b T$, and I is the ionic

strength. The dependence of U^{Coul} and U^{GB} on λ is realized through linear scaling of the partial charges on titratable residues between the protonated and the unprotonated forms. In an analogous fashion, the van der Waals interactions involving titratable hydrogens are also attenuated by the titration coordinates. The last term in eq 1 represents the biasing energy acting on the titration coordinates:

$$U^*(\{\boldsymbol{\theta}_i\}) = \sum_i -U^{mod}(\boldsymbol{\theta}_i) + U^{pH}(\boldsymbol{\theta}_i) + U^{barr}(\boldsymbol{\theta}_i) \quad (5)$$

where U^{mod} is a potential of mean force function for deprotonating a model compound in solution along the titration coordinate, U^{barr} is a harmonic potential which suppresses the residence time of unphysical intermediate values of λ, and U^{pH} provides the additional free energy for the protonation equilibrium due to solution pH:

$$U^{pH}(\lambda_i) = \log(10)k_bT(pK_a^{mod} - pH)\lambda_i \quad (6)$$

where pK_a^{mod} is the experimentally determined pK_a of a model compound in solution.

Continuous Constant pH Molecular Dynamics in Explicit Solvent. The key to CPHMD and other continuous titration methods is to simultaneously derive forces on the spatial and titration coordinates. While it is straightforward to compute forces on spatial coordinates in explicit-solvent simulations, there is inherent difficulty in the latter due to the need for very accurate estimate of the electrostatic desolvation free energy (see U^{GB} in eq 3). In fact, attempts to directly calculate the free energy of charging titratable residues repeatedly during molecular dynamics by considering explicit interactions between solvent molecules and solute have encountered severe convergence problems in the context of both discrete[7] and continuous constant pH MD methods.[6,30] Our own tests revealed that the variance in the instantaneous forces on the titration coordinates is up to an order of 100 kcal/mol per λ unit, whereas the forces exerted from the pH biasing energy 1 pH unit away from the model compound pK_a is only 1.3 kcal/mol per λ unit. Therefore, we decided to use a "mixed-solvent" scheme, where the GB model is used to derive forces on the titration coordinates, while the explicit-solvent model is used to propagate the spatial coordinates. To enable a direct coupling between solvent dynamics and proton titration of solute, we retain the λ-dependent scaling of van der Waals interactions involving titrating hydrogens and solvent molecules. An analogous "mixed solvent" scheme has been developed by Baptista and co-workers and applied in the context of the discrete constant pH MD for protein titration studies.[8] One important difference is that their scheme does not include a direct (van der Waals) coupling between solvent dynamics and solute titration.

The caveat of the "mixed solvent" scheme is that no formal Hamiltonian exists and that potential artifacts may occur. Since the solvation-related force on titration coordinates is treated in a mean-field manner without explicitly accounting for the electrostatic interactions with nearby water molecules, inadequate or lagged response of solvent to the change in the charge state of the titrating site may occur. We expect this undesirable side effect to be minimal because of the aforementioned van der Waals coupling between solute protonation and solvent dynamics, and because in continuous evolution of titration coordinates, the energy change is small at each time step. Nevertheless, a preventive fix is to increase the time step for λ coordinates (currently the same as spatial coordinates), thereby allowing relaxation of surrounding solvent molecules. Such a strategy has been demonstrated to be very effective in the discrete constant pH molecular dynamics simulations using the "mixed-solvent" scheme.[8] Another source for potential artifacts in this and other "mixed-solvent" simulations is related to the fact that the total energy is no longer strictly conserved, which may result in a drift or pronounced fluctuation in temperature and energy of the sytem. We will examine these potential artifacts later in detail.

pH-Based Replica-Exchange Sampling Protocol. It has been noted previously[9,13,10] that in constant pH molecular dynamics, the convergence of protonation-state sampling and resulting pK_a values is slow due to the tight coupling of conformational dynamics and protonation equilibria. To address this issue, the temperature-based replica-exchange (T-REX) protocol[31] was applied to enhance conformational sampling in the GB-based continuous[26] and discrete[32] constant pH methods, which has led to significant improvement in the convergence of calculated pK_a values. A straightforward implementation of the T-REX protocol in explicit-solvent simulations is however not effective because of the large number of replicas needed to account for the solvent degrees of freedom.[33] Recently, Simmerling and co-workers have proposed a mixed-solvent scheme to reduce the number of replicas,[34] which may be incorporated into the explicit-solvent CPHMD presented in this work. One issue that was noted,[34] and is currently being addressed,[19] is the distorted conformational distribution due to inaccuracy of the underlying implicit-solvent model. To avoid this problem, we decided to enhance the sampling of protonation space directly by making use of a REX protocol based on the pH-biasing energy (eq 6). This protocol is a specific application of the reaction-coordinate replica-exchange method.[35]

In the pH-based REX protocol, simulations of independent replicas are run at the same temperature but different pH conditions. The exchange of pH conditions between a pair of replicas adjacent in pH is periodically attempted according to the Metropolis criterion, which gives the exchange probability as

$$P = \begin{cases} 1 & \text{if}\Delta \le 0 \\ \exp(-\Delta) & \text{otherwise,} \end{cases} \quad (7)$$

where Δ represents the change in the total pH-biasing energy defined as

$$\Delta = \beta(U^{pH}(\{\boldsymbol{\theta}_i\};pH') + U^{pH}(\{\boldsymbol{\theta}_i'\};pH) - U^{pH}(\{\boldsymbol{\theta}_i\};pH) - U^{pH}(\{\boldsymbol{\theta}_i'\};pH')) \quad (8)$$

Here β is the inverse temperature, the first two terms are the pH-biasing potential energies (eq 6) for the first and second replicas after the exchange, and the last two terms are the corresponding energies before the exchange.

SIMULATION DETAILS

Model Compounds. As in the previous work,[13,26] model compounds for Asp, Glu, His, and Lys side chains are single amino acids acetylated at N-terminus (ACE), and N-methylamidated at C-terminus (CT3). The model pK_a values (used in eq 6) were 4.0, 4.4, and 10.4 for Asp, Glu, and Lys, respectively.[36] The model pK_a of His was taken as 6.6 and 7.0 for the Nδ and Nϵ sites, respectively.[37] The model compound for the C-terminus attached to phenylalanine (CT-Phe) in HP36 was the acetylated

C-terminal hexapeptide (KEKGLF) from HP36 with a measured pK_a of 3.2.[38] The parameters in the potential of mean force function U^{mod} were determined using thermodynamic integration (TI) in explicit solvent.[13] Parameterization simulations at each combination of λ and x for double-site titratable residues were run for 1 ns. In the TI procedure, the protonation states of other titratable residues in the model peptide for CT-Phe were fixed because their pK_a's are at least 1 pH unit higher than the C-terminus. Except for CT-Phe, the ionic strength in the GB calculation was set to zero during the TI simulations following the previous protocol.[13] For CT-Phe the ionic strength was 150 mM in accord with experiment.[38]

Proteins. Five proteins were studied in this work: the 45-residue binding domain of 2-oxoglutarate dehydrogenase multienzyme complex, BBL (PDB: 1W4H), the 36-residue subdomain of villin headpiece, HP36 (PDB: 1VII), the 56-residue N-terminal domain of ribosomal L9 protein, NTL9 (PDB: 1CQU), the 149-residue, of which 129 residues were resolved in the crystal structure, hyperstable variant of staphyloccal nuclease Δ+PHS, SNase (PDB: 3BDC), and the 129-residue hen egg white lysozyme, HEWL (PDB: 2LZT). For all structures, the HBUILD facility of CHARMM[39] was used to add hydrogens. Unless otherwise specified, no explicit ions were added in the pH-REX simulation because of the small simulation box and the low ionic strengths used in experiment. See later discussions. The ionic strengths in the GB calculations were set to 200, 150, 100, 100, and 50 mM for BBL, HP36, NTL9, SNASE, and HEWL, respectively, consistent with the experimental conditions.[38,40−43] Unless otherwise noted, both N- and C-termini of proteins were left in the free, charged form. For SNase, the published crystal structure was missing residues 1−6 and 142−149. To avoid potential errors, the structure was acetylated at N-terminus and amidated at C-terminus. For NTL9, the C-terminus was amidated in accord with experiment.[41]

Simulation Protocol. We have implemented the explicit-solvent CPHMD method in a developmental version of CHARMM (c35b3)[39] and the pH-REX sampling scheme in the MMTSB tool set.[44] All of the simulations described in this work were performed with the all-atom CHARMM22/CMAP force field for proteins[45] and the CHARMM modified version of the TIP3P water model.[46] The solvation forces on the titration coordinates were calculated using the GBSW implicit-solvent model[28] with the refined[18] atomic input radii of Nina et al.[47] The SHAKE algorithm was applied to all hydrogen bonds and angles to allow a 2 fs time step. Nonbonded electrostatic interactions were calculated using the particle-mesh Ewald summation with a charge correction to reduce pressure and energy artifacts for systems with a net charge.[48] In the GB calculation, all input parameters were identical to the previous work.[26]

All simulations were performed under ambient pressure and temperature conditions using the Hoover thermostat[49] with Langevin piston pressure coupling algorithm.[50] Proteins and model compounds were built and then placed in a truncated octahedron water box of a specified size such that the distance between the solute and the edges of the box was at least 14 Å. Water molecules within 2.6 Å of any heavy atom of the solute were deleted. Energy minimization was carried out in three stages. First, a harmonic restraint with a force constant of 50 kcal/mol·Å was applied to solute heavy atoms, and the structure was energy minimized with 50 steps of the steepest descent (SD) and 200 steps of the adoptive basis Newton−Ralphson (ABNR) methods. Then the force constant was reduced to 25 kcal/mol·Å, and the same minimization protocol was applied. Finally, the force constant was reduced to 10 kcal/mol·Å, and the structure was energy minimized with 5 SD and 20 ABNR steps.

In the pH-REX simulation of a model compound, three pH replicas, one at the reference pK_a and two at 1 pH unit above and below the reference value, were used. Three independent pH-REX simulations were conducted, where each REX simulation lasted 1.2 ns per replica and the first 200 ps was discarded in the pK_a calculation. For proteins, one pH-REX simulation was performed. In the pH-REX protocol, the pH spacing was 1 pH unit, and the pH range extended at least 1 unit above and below the highest and lowest experimentally determined pK_a value for the protein. Specifically, for BBL the pH range is 2−9, for HP36, NTL9, and SNase it is 0−7, and for HEWL it is 0−9. Each pH replica was subjected to 4 ps of restrained equilibration without pH exchange, where a harmonic potential with the force constant of 10 kcal/mol·Å was applied to all solute heavy atoms. Following equilibration, unrestrained simulation with the pH-REX protocol was performed. The exchange in pH was attempted every 100 dynamic steps or 0.2 ps for model compound and 500 steps or 1 ps for protein simulations. The success rate for exchanges was at least 40%. Protein simulations lasted 2 ns, and the first 0.25 ns was discarded in the analysis and the pK_a calculation. Simulation of HP36 was run for 4 ns in order to observe pK_a behavior at longer simulation times.

Calculation of pK_a Values. To calculate the pK_a of a titratable site, we first recorded the population of protonated ($\lambda < 0.1$, N^{prot}) and unprotonated ($\lambda > 0.9$, N^{unprot}) states from simulations of different pH replicas. The resulting unprotonated fractions S at multiple pH values were then fitted to the following modified Hill equation, in accord with the commonly used model for fitting pH-dependent NMR chemical shifts:[42]

$$S(\mathrm{pH}) = \frac{s_{A^-} + s_{HA}10^{n(\mathrm{p}K_a - \mathrm{pH})}}{1 + 10^{n(\mathrm{p}K_a - \mathrm{pH})}} \quad (9)$$

where n is the Hill coefficient, which represents the slope of the transition region of the titration curve,[42] s_{A^-} and s_{HA} are fitting parameters, which represent the extrapolated S values at extreme acidic and basic pH conditions for the observed titration event. Equation 9 becomes the Hill equation when protonation or unprotonation is complete in the simulated pH range, e.g., $s_{A^-} = 1$ and $s_{HA} = 0$, which was the case for nearly all residues. Occasionally, for acidic residues with significant negative pK_a shifts, s_{HA} deviated significantly from 0 as a result of incomplete protonation at the lowest pH condition. Finally, to account for the small systematic deviations of calculated pK_a's of model compounds relative to the reference values, we made the following postcorrections: Asp (+0.2), Glu (+0.3), and His (−0.3) to the pK_a values of proteins.

RESULTS AND DISCUSSION

Stability of Trajectory. Before applying the explicit-solvent CPHMD to titration simulations, it is important to examine potential artifacts due to caveats in the mixed scheme and the change in total net charge. As mentioned earlier, the proposed method does not conserve energy because the protonation states of titratable groups are changed using an implicit description of the electrostatic interactions with solvent, which may lead to drift or increased fluctuation in temperature and energy of the simulated system. Another source for potential artifacts is related

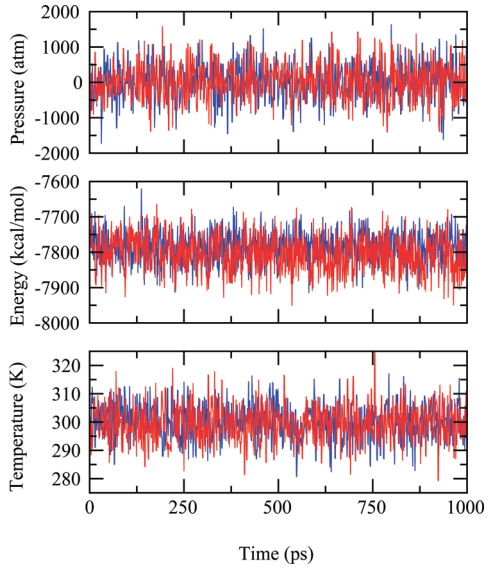

Figure 1. Instantaneous pressure, potential energy, and temperature in the explicit-solvent CPHMD simulation of lysine at pH of 10.4 (blue) and in the simulation of the neutral lysine with fixed protonation state (red).

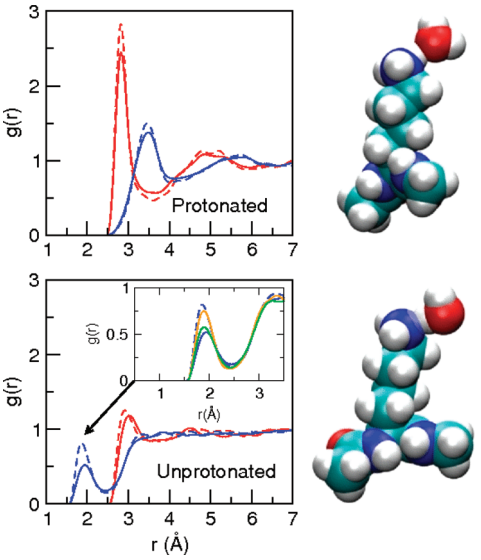

Figure 2. Response of explicit solvent to titration. Radial distribution function for the titratable nitrogen atom of lysine to the hydrogen (blue) or oxygen (red) atom of water. Dashed lines are from the simulation with the fixed protonation state; solid lines are from the CPHMD titration with protonated (charged) and deprotonated (neutral) states coexisting. Snapshots of the interacting water and lysine are shown. The charged lysine donates a hydrogen bond to water (upper), while the neutral lysine accepts a hydrogen bond from water (lower). Simulations with fixed protonation states were run for 1 ns. The CPHMD titration time was 2 ns, and the deprotonated fraction was about 0.5. The inset gives radial distribution functions when a very stringent cutoff ($\lambda > 0.99$) is used to define the deprotonated state (green) and when λ values are updated every 10 MD steps in addition to the stringent cutoff (orange).

to the fluctuating net charge on the system during proton titration. In the default implementation of Ewald summation a neutralizing plasma, which is a uniform distribution of a charge equal and opposite to the net charge, is added to the summation to avoid divergence in Coulomb energy for periodic systems.[51] This background plasma has been noted to introduce pressure artifacts for small net-charged systems, which could dramatically affect the dynamics of simulations at constant pressure.[48] Brooks and co-workers showed that the artifacts are drastically reduced by invoking a charge correction term.[48] We applied this correction term in all of our simulations.

To assess the extent of the spurious effects, we examined the temperature, pressure, and total potential energy of the system along the trajectory using two protocols. In the first protocol, a blocked lysine was subjected to CPHMD titration at pH of 10.4. In the second protocol, a fixed-charge simulation was conducted using a neutral lysine with an otherwise identical simulation setup. As shown in Figure 1, the time series for temperature, pressure, and potential energy in the CPHMD titration of lysine (with 1:1 protonated and deprotonated states) is virtually indistinguishable from the conventional simulation of neutral lysine with fixed protonation state. The pressure fluctuations are quite large for both systems, but this is expected because of the small size of the simulation box. Also, any energy leaking into or out of the system due to the nonconservative change in protonation state is not readily apparent, as there is no visible drift in the total potential energy for this system. To further verify the stability of pressure, temperature, and potential energy, we performed CPHMD titrations for other model compounds and proteins. No systematic drift or increased fluctuation was observed in any of the three quantities at the simulation time scales (several nanoseconds) for either model compounds or proteins. Thus, we conclude that, with the net charge correction and the Hoover thermostat, potential artifacts in pressure, temperature, and potential energy are negligible.

Response of Explicit Solvent to Titration. Although the van der Waals interactions between titratable hydrogen atoms and solvent molecules are explicitly described, the lack of explicit treatment of electrostatic interactions may have an undesirable effect such that water molecules cannot adjust quickly to a low-energy position following a change in the titration coordinate, resulting in an unrealistic arrangement of solvent around the titrating site. To examine the response of explicit water molecules to solute titration, we calculated the radial distribution functions (RDF) for the charged (protonated) and neutral (unprotonated) lysine from the (conventional) simulations (one for charged and one for neutral) and compared them with the RDF's from one CPHMD titration simulation. The latter simulation was conducted at a pH condition such that the charged and neutral populations are almost equal. As seen in Figure 2, the positions of maxima and minima in the RDF's of the charged and neutral forms of Lys are identical in the conventional simulations and the CPHMD titration, which demonstrates that the water structure is qualitatively indistinguishable. To further investigate the reorientation of water molecules in response to titration, we took a closer look at the solute−solvent interactions that give rise to the peaks of the RDF's. Interestingly and reassuringly, the relative orientation of lysine and the nearby water is identical in the conventional simulations and the CPHMD titration. Figure 2 also shows the representative snapshots of the charged and neutral lysines interacting with an adjacent water molecule. When lysine is charged, it acts as a hydrogen-bond donor, interacting with the oxygen atom of water. When lysine is neutral, it acts as a hydrogen-bond acceptor, interacting with the hydrogen atom of water.

Despite the remarkable agreement in the positions of maxima and minima of the RDF's, the amplitude of the peaks from the

Table 1. Calculated and Experimental pK_a Values of Model Compounds

residue	calcd[a]	calcd[b]	ref[c]	pace lab[d]
Asp	3.79 ± 0.09	3.77 ± 0.02	4.0	3.67 ± 0.04
Glu	4.09 ± 0.11	4.05 ± 0.01	4.4	4.25 ± 0.05
His	6.89 ± 0.08	6.89 ± 0.01	6.6/7.0	6.54 ± 0.04
Lys	10.21 ± 0.02	10.41 ± 0.02	10.4	10.40 ± 0.08
CT-Phe	3.38 ± 0.06		3.2[e]	–

[a] Results using the standard simulation protocol where the λ value was updated every MD step and the simulation length was 1.2 ns per pH replica. The average pK_a's obtained by fitting S data from three independent pH-REX titrations are listed along with one-half of the difference between the highest and lowest calculated values. [b] Results from test simulations where the λ value was updated every 10 MD steps and the simulaton length was 10 ns per pH replica. [c] Measured pK_a's based on the blocked single amino acids from Nozaki and Tanford.[36] These model pK_a's were used in the pH biasing energy (eq 6). For His, the listed pK_a's are the microscopic values for δ and ϵ sites. The resulting macroscopic pK_a is 6.45.[13] Errors in the measurements are typically ±0.1−0.2.[57] [d] The most recent data from Pace lab based on potentiometric titrations of alanine pentapeptide Ac-AA-X-AA-NH_2 , where X denotes the titrating residue.[57] [e] Measured pK_a of the C-terminal carboxylic acid in the C-terminal peptide of HP36 (sequence KEKGLF) based on the NMR titration data from Raleigh lab.[38]

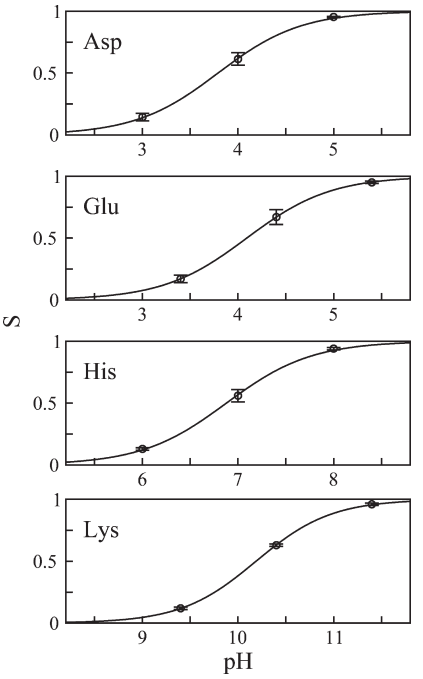

Figure 3. Titration curves for the blocked model compounds from the explicit-solvent CPHMD simulations. Three independent pH-REX simulations were performed. Each REX simulation utilized three pH replicas with each replica running for 1 ns. The average unprotonated fractions S (calculated from the three runs and shown as circles) at three pH values were fit to the Hendenson−Hasselbach equation and shown as lines. At each pH, an error bar indicates the range of the calculated S values, which is the largest at the pH closest to the pK_a value. These ranges are 0.10, 0.12, 0.10, and 0.02 for Asp, Glu, His, and Lys, respectively.

CPHMD titration is reduced as compared to those from conventional simulations. This reduction in the amplitude of RDF can be mainly attributed to the slight lagging in water equilibration following a switch in protonation state and to a lesser extent the cutoff chosen in our definition of protonated and deprotonated states. The inset in Figure 2 shows that with a very stringent cutoff (λ > 0.99) there is small improvement in the amplitude of the RDF. If we use the stringent cutoff combined with a λ update of every 10 MD steps, the amplitude of the RDF is dramatically increased to nearly superimpose on the result from the simulation with fixed protonation state. If the frequency of switching protonation state is much slower, then the RDF's would exactly match those calculated from the simulations at fixed charge. Baptista and co-workers showed that in the MD simulation, the reorganization time of water following the most dramatic protonation event from the fully neutral to doubly charged state of succinic acid is 1−3 ps.[8] Considering the average residence time at either protonation state in our simulation was on average about 1 ps and the transition between protonation states is continuous, water molecules have sufficient time to rotate to a favorable position following titration. Nevertheless, the data of lysine titration shows that the update frequency or time step for propagation of titration coordinates (currently set to be the same as the propagation of conformational dynamics) can be increased to ensure the full extent of water relaxation. A potential drawback is the slow down of protonation-state sampling.

Convergence and Accuracy of Model Compound Titrations. Before attempting to perform titration simulations of proteins, it is important to assess the required simulation time to reach converged values for the unprotonated fraction (S) of model compounds as well as the accuracy and the precision of the calculated pK_a's. We first examine titration simulations conducted at a single pH value. Explicit-solvent CPHMD titration of a blocked lysine was performed at the pH equal to the reference pK_a of 10.4. The S values stabilized at about 5 ns, and there was little change over the remainder of the 10 ns simulation. We repeated the simulation twice with different randomly assigned velocities and observed a similar convergence time. Similar results were also found for the blocked Asp, Glu, and His which have two titration sites. The lengthy simulation time (5 ns) required for the convergence of pK_a values for single amino acids indicates the need for accelerated sampling. To directly enhance the protonation-state sampling, we applied the pH-based replica-exchange protocol with three replicas placing at pH values of 9.4, 10.4, and 11.4 in the lysine titration. The S values were converged within 1 ns for all model compounds, demonstrating significant acceleration over the single pH simulation. We summarize these results in Table 1. The uncertainty or random error in the calculated model compound pK_a's ranges from 0.02 to 0.11, which is similar to the range found in potentiometric and NMR titration experiments (see Table 1). To further assess convergence, we examine the reproducibility of S values and the quality of fitting to the Henderson−Hasselbach equation. In Figure 3 results of three independent pH-REX simulations (1 ns/replica) for Asp, Glu, His, and Lys are shown. The error in the S value ranges from 0.02 to 0.12, and the χ-square value of the fitting is virtually 0. Thus, the above data demonstrate that 1 ns pH-REX titrations offer converged sampling for protonation equilibria.

Next we examine the accuracy of the calculated pK_a's of model compounds. As compared to the target reference values, the pK_a's of Asp, Glu and Lys are underestimated by 0.2−0.3 pH units, while that of His is overestimated by 0.3 pH units

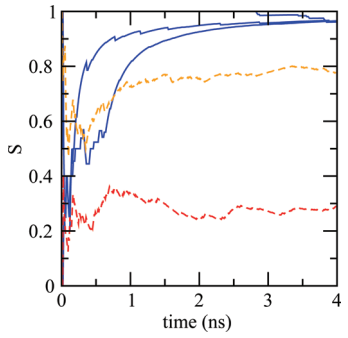

Figure 4. Enhancement of protonation-state and conformational sampling for protein titrations. Cumulative unprotonated fraction of Asp44 of HP36. Data from the pH-REX simulations are shown in red for replica at pH of 2 and orange for replica at pH of 3. Data from three independent single pH simulations at pH 2.3 are shown in blue.

(Table 1). There are two possible sources for the systematic deviations. The first possibility has to do with artifacts in simulations of net charged periodic systems using Ewald potential. Even with the net charge correction, Brooks and co-workers noted that the charged form may be slightly favored in the free energy simulation of a single ion, and this deviation depends on the size of the simulation box.[48] Our tests however showed that increasing the box size did not affect the pK_a results for model compounds. We further ruled out the net charge related artifact because the same systematic errors, e.g., underestimation of the pK_a's for Asp and Glu and overestimation of the pK_a for His, were also observed in the GB-based CPHMD simulations.[26]

The systematic errors in pK_a's indicate that the deprotonation free energy based on the potential of mean force function, which is determined by the thermodynamic integration (TI) procedure, does not exactly match that in the titration simulation. One possible reason for the discrepancy is the difference in water relaxation because in the TI simulation water has more time to relax at a specific λ value than in the titration simulation. To investigate this issue, we repeated the titrations with slower λ dynamics, updating λ every 10 MD steps. Interestingly, the deviation for the pK_a of Lys is abolished, but the deviation for Asp, Glu and His remained. Examination of the λ and x trajectories revealed that the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His) occasionally experience prolonged residence time. In the absence of extensive analysis and consideration, we suggest that one route for correcting this bias is to make the barrier in the x (tautomeric) dimension a function of λ such that when λ approaches the degenerate protonation state, interconversion becomes increasingly difficult. This is clearly a limitation that needs to be addressed in our future work. Nevertheless, since this bias is present in both model compound and protein titrations, the effect on the calculated pK_a shifts is negligible. To correct for the systematic deviations, we added post corrections for all the calculated pK_a values of proteins (see Simulation Details Section).

Enhanced Sampling of Protonation and Conformational States of Proteins. We have demonstrated that the pH-REX protocol significantly accelerates the pK_a convergence for model compounds. Now we show that the pH-REX protocol significantly enhances sampling in both protonation and conformational space for proteins. Take the titration of HP36 as an example. Figure 4 displays the time series of the unprotonated

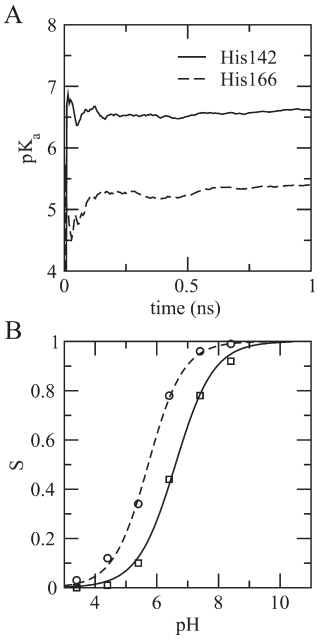

Figure 5. Convergence of protein titrations. (A) Time series of the calculated pK_a's for BBL from the explicit-solvent CPHMD simulations with pH-REX protocol. The S values at pH of 7 and 6 are used for His142 and His166, respectively. (B) Titration data based on the 1 ns simulation and best fits to the modified HH equation (eq 9).

fraction for Asp44 from one pH-REX simulation and three single pH simulations. In the single pH simulations, Asp44 was trapped in the unprotonated form at pH 2.3, as a result of the persistent salt-bridge interaction with Arg15. In the pH-REX simulation, however, both protonated and unprotonated forms of Asp44 were sampled at pH 2 and 3, because the simulation was able to capture both formation and disruption of the salt bridge. Thus, by making use of the direct coupling between protonation events and conformational dynamics, the pH-REX protocol allows the protein to overcome local energy barriers, while retaining the correct thermodynamic distribution. In this regard, pH-REX has a similar effect as the temperature-based REX protocol, which significantly accelerates the sampling convergence of both protonation and conformational states in the GB-based CPHMD simulations.[26]

Convergence and Overall Accuracy of Protein Titrations. In order for titration simulations to be practical, protonation-state sampling needs to converge within a reasonable amount of time. While we have shown that 1 ns of pH-REX titration is sufficient for obtaining converged pK_a's for model compounds, we also observed that 1 ns titration also yields converged pK_a's for proteins, despite the fact that the degrees of freedom in a protein system may be orders of magnitude greater as compared to a model compound. This seemingly surprising observation is consistent with data from the GB-based CPHMD simulations[26,12] and can be attributed to the fact that pK_a's are mainly determined by local environment. To illustrate the rapid convergence in protein titrations, we monitor the times series of the S value and pK_a as well as the quality of fitting. In Figure 4 we can see that the S values for HP36 stabilize at 1 ns. The small fluctuation after 1 ns does not cause noticeable change in the pK_a value because of the logarithmic relationship between S and pK_a. Figure 5A shows that, after only a few hundred ps, the calculated pK_a's of the two histidines in BBL become stable and do not change in the remaining simulation

Table 2. Calculated and Experimental pK_a Values in HP36, BBL, and NTL9

residue	explicit solvent[b]	GB	expt[a]
	BBL		
His142	6.94 ± 0.06 (6.83)	6.47 ± 0.03	6.47 ± 0.04
His166	5.78 ± 0.04 (5.90)	4.84 ± 0.19	5.39 ± 0.02
	HP36		
Asp44	2.66 ± 0.09 (2.77)	3.17 ± 0.11	3.10 ± 0.01
Glu45	3.36 ± 0.31 (3.28)	3.49 ± 0.09	3.95 ± 0.01
Asp46	3.03 ± 0.09 (3.12)	3.51 ± 0.03	3.45 ± 0.12
Glu72	3.50 ± 0.21 (3.45)	3.53 ± 0.10	4.37 ± 0.03
CT-Phe	3.31 ± 0.20 (3.16)	3.16 ± 0.14	3.09 ± 0.01
			3.24 ± 0.12
	NTL9		
Asp8	2.83 ± 0.07 (2.80)	3.19 ± 0.20	2.99 ± 0.05
Glu17	3.57 ± 0.14 (3.50)	3.67 ± 0.13	3.57 ± 0.05
Asp23	2.75 ± 0.16 (2.82)	2.11 ± 0.11	3.05 ± 0.04
Glu38	3.38 ± 0.30 (3.40)	3.70 ± 0.19	4.04 ± 0.05
Glu48	3.47 ± 0.17 (3.42)	3.74 ± 0.20	4.21 ± 0.08
Glu54	3.65 ± 0.22 (3.49)	3.64 ± 0.08	4.21 ± 0.08
avg abs dev	0.44 (0.45)	0.36	
rmsd	0.50 (0.52)	0.47	
max abs dev	0.87 (0.92)	0.99	

[a] pK_a's determined by NMR titration for BBL,[40] HP36,[38] and NTL9.[41] [b] Values in parentheses were obtained from the 2 ns simulation.

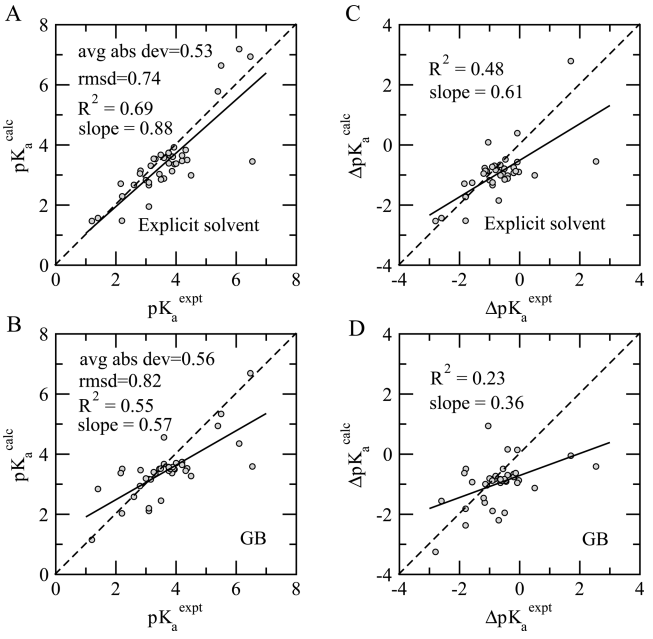

Figure 6. Comparison between calculated and experimental pK_a values and pK_a shifts relative to model values. Calculated pK_a values from the explicit-solvent and GB-based titrations are shown in A and B, respectively. Calculated pK_a shift from explicit-solvent and GB-based titrations is shown in C and D, respectively. Regression line (solid), slope, and R^2 value are shown on each plot as wells as $y = x$ line (dashed) to facilitate visual comparison.

time. This is encouraging given the fact that one of the histidines is buried and as such may require more sampling. Another indication of convergence is the quality of fitting to the HH equation. Figure 5B shows nearly perfect fits ($R^2 > 0.95$) for both residues based on the 1 ns titration data.

To assess the overall accuracy of the explicit-solvent CPHMD method, we performed titration on five test proteins, HP36, BBL, NTL9, SNase, and HEWL and compared the calculated pK_a's with experiment as well as the GB-based simulations, where the latter used the same pH-REX protocol and salt as well as temperature conditions. The results are presented in Tables 2, 4, and 5 along with the estimates of statistical uncertainty, which were calculated as half of the difference between the pK_a's calculated from the first and last half of the 750 ps simulation. The total simulation length was 1 ns, and the data from the first 250 ps were discarded. As a validation of convergence, the pK_a's calculated using 2 ns simulations are also listed. In reference to experimental data, the overall root-mean-squared deviation (rmsd) from the explicit-solvent titrations is 0.74, which is slightly lower than the rmsd from the GB-based titrations (0.82). As a more informative measure of calculation accuracy, linear regressions of the calculated vs measured pK_a shifts are shown in Figure 6 for the explicit-solvent and GB simulations. While the R^2 value and the slope are 0.48 and 0.61, respectively, from the explicit-solvent titrations, they are 0.23 and 0.36 from the GB titrations. Since the correlations are relatively low, we repeated the regression analysis by removing the data points with the four largest absolute pK_a shifts. The R^2 value from the explicit-solvent simulations dropped from 0.48 to 0.25, while R^2 from the GB simulations also dropped dramatically, from 0.23 to 0.06. Thus, the results show that the improvement due to explicit solvent is robust. Since the data set comprised of mainly acidic residues, the slopes being below 1 suggest that both simulations overestimate the negative pK_a shifts or underestimate the pK_a's. A close examination of the correlations reveals that the significantly improved agreement with experiment in the explicit-solvent titrations is due to reduction of relatively large errors for several groups. Thus, overall the explicit-solvent simulations offer an increased accuracy for predicting protein pK_a's. The reasons in specific cases will be delineated next.

Small Proteins BBL, HP36, and NTL9. We first examine the performance of the explicit-solvent CPHMD titrations for three small proteins with 36−56 residues and all α as well as mixed α−β topologies. The results are listed in Table 2 along with the GB titration data. The convergence of both explicit- and implicit-solvent titrations is excellent. The largest difference between the pK_a's calculated from the first and the last half of the simulation is 0.3 units. Extending the explicit-solvent simulations to 2 ns leads to a maximum pK_a change of 0.15 units and does not improve the agreement with experiment. Overall, the explicit-solvent data are similar to the GB data. The rms as well average absolute and maximum deviations from experiment in the explicit-solvent titration are 0.50, 0.44 and 0.87, respectively, similar to the GB titration. The deviations from experiment arise from the overestimation of the negative pK_a shifts of acidic residues in both explicit- and implicit-solvent titrations.

We examine two cases where the pK_a's from the explicit-solvent titration are at least 0.6 pH units different from the GB titration. In both cases, the explicit-solvent titration improves agreement with experiment. Asp23 is a residue where the explicit-solvent titration reduces the overestimation of the pK_a downshift of Asp23 from 0.9 to 0.3 units. This is because the salt-bridge

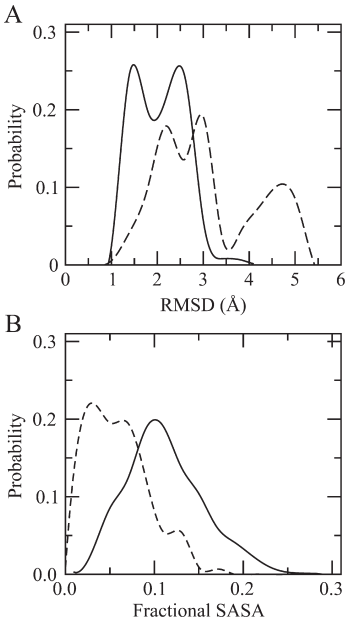

Figure 7. Structural comparison of BBL from explicit-solvent (solid) and GB (dashed) simulations at pH 5. (A) Probability distributions of backbone rmsd. (B) Ratio of the solvent accessible surface area (SASA) of His166 in BBL relative to the solvent-exposed value.

interaction with the nearby amino terminus was overstabilized in the GB simulation, a known problem in GB models.[13] His166 is the only buried residue in this data set. While being excluded from solvent, it also interacts with three nearby lysines. Thus, both desolvation and electrostatic repulsion destabilize the protonated or charged form of His166, leading to a downward pK_a shift relative to the model value. This is reflected in the experimental pK_a of 5.39, about 1.1 pH units lower than the model value. In the explicit-solvent titration the pK_a shift is underestimated by 0.41 pH units, while it is overestimated in the GB titration by 0.55 pH units. Detailed analysis of the trajectories reveals the major cause of the difference to be structural. Figure 7A shows that in the explicit-solvent simulation, the conformations stayed close to the starting structure with the backbone rmsd centered at 2 Å. In the GB simulation, however, a conformational cluster developed that significantly deviates from the initial structure with the backbone rmsd centered at 4.9 Å. Figure 7B shows that while His166 is slightly exposed to solvent in the explicit-solvent simulation, it is fully enclosed in the GB simulation. Examination of the average distances to the nearby lysines reveals that the Coulomb interactions in both explicit-solvent and GB simulations are similar. Therefore, we suggest that the overestimation of the pK_a shift for His166 in the GB simulation is mainly due to the overestimation of desolvation penalty as a result of exaggerated cloistering of His166. Reduced mobility, especially of buried sites, has been also observed in other GB simulations.[24]

Although for these small proteins the explicit-solvent pK_a calculations are quite accurate, it is important to further discuss another issue concerning the explicit-solvent CPHMD method. Since the net charge is changing and may become large depending on the protonation state of the protein, we examined the effect of adding an approximate number of counterions to minimize the net charge of the system in all pH conditions. Because of the large number of basic residues of NTL9 and the resulting net positive charge, NTL9 is an ideal test case to quantify the magnitude of the effect. As shown in Table 3 the calculated pK_a values in the simulations with neutralizing counterions are virtually identical to those where no net charge neutralizing ions were added. Thus, at least for the short simulation time required to obtain converged pK_a values, the data indicate that it is not necessary to include neutralizing ions.

Table 3. Effects of Adding Explicit Ions on Calculated pK_a Values in NTL9

residue	calcd[b]	ions[c]	expt[a]
Asp8	2.83 ± 0.07	2.91 ± 0.31	2.99 ± 0.05
Glu17	3.57 ± 0.14	3.38 ± 0.19	3.57 ± 0.05
Asp23	2.75 ± 0.16	2.98 ± 0.16	3.05 ± 0.04
Glu38	3.38 ± 0.30	3.48 ± 0.04	4.04 ± 0.05
Glu48	3.47 ± 0.17	3.42 ± 0.34	4.21 ± 0.08
Glu54	3.65 ± 0.22	3.52 ± 0.25	4.21 ± 0.08
avg abs dev	0.41	0.40	
rmsd	0.48	0.49	
max abs dev	0.73	0.78	

[a] pK_a's determined by NMR titration.[41] [b] Calculated pK_a's from explicit-solvent titrations without counterions (as listed in Table 2). [c] Calculated pK_a's from simulations with an identical set up except for the addition of Cl^- ions such that the net charge of the protein at all pH conditions was minimimzed.

Table 4. Calculated and Experimental pK_a Values in SNase

residue	explicit solvent[c]	GB	expt[a]
Glu10	3.14 ± 0.09 (3.33)	3.47 ± 0.01	2.82 ± 0.07
Asp19	2.29 ± 0.15 (2.49)	3.51 ± 0.02	2.21 ± 0.07[b]
			6.54 ± 0.06
Asp21	3.45 ± 0.28 (3.55)	3.59 ± 0.00	3.01 ± 0.01
			6.54 ± 0.02[b]
Asp40	3.13 ± 0.23 (3.35)	3.37 ± 0.09	3.87 ± 0.09
Glu43	3.83 ± 0.08 (3.76)	3.45 ± 0.00	4.32 ± 0.04
Glu52	3.92 ± 0.01 (3.88)	3.52 ± 0.02	3.93 ± 0.08
Glu57	3.67 ± 0.16 (3.64)	3.52 ± 0.01	3.49 ± 0.09
Glu67	3.66 ± 0.06 (3.67)	3.45 ± 0.06	3.76 ± 0.07
Glu73	3.53 ± 0.11 (3.54)	3.36 ± 0.13	3.31 ± 0.01
Glu75	3.54 ± 0.27 (3.58)	3.40 ± 0.06	3.26 ± 0.05
Asp77	<0.0 (<0.0)	3.14 ± 0.03	<2.2
Asp83	2.54 ± 0.12 (2.84)	3.50 ± 0.04	<2.2
Asp95	2.71 ± 0.57 (2.97)	3.37 ± 0.06	2.16 ± 0.07
Glu101	3.64 ± 0.11 (3.67)	3.51 ± 0.01	3.81 ± 0.10
Glu122	3.61 ± 0.03 (3.75)	3.57 ± 0.01	3.89 ± 0.09
Glu129	3.74 ± 0.11 (3.71)	3.57 ± 0.12	3.75 ± 0.09
Glu135	3.39 ± 0.20 (3.44)	3.56 ± 0.03	3.76 ± 0.08
avg abs dev	0.46 (0.48)	0.63	
rmsd	0.86 (0.85)	0.96	
max abs dev	3.09 (3.00)	2.95	

[a] pK_a determined by NMR titration.[42] [b] The major transition when the experimental data was fit to a two pK_a model. [c] Values in parentheses were obtained from the 2 ns simulation.

SNase. The calculated pK_a's for a larger protein, a hyperstable variant of the 149 residue SNase, are summarized in Table 4.

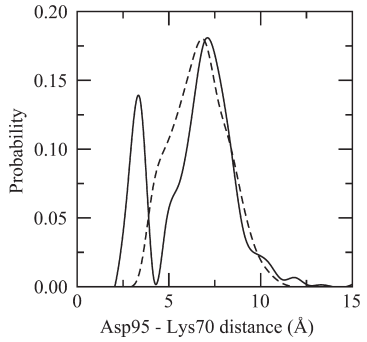

Figure 8. Probability distribution of the minimum distance between the carboxylate oxygens of Asp95 and amino nitrogen of Lys70 of SNase from the explicit-solvent (solid) and GB (dashed) titrations at pH of 3.

SNase is a good test system because the structure-based continuum calculations gave very poor agreement with experiment presumably due to the lack of explicit treatment of protein flexibility.[42] Overall, the explicit-solvent titration offers a better agreement with experiment. The rms as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.86, 0.46, and 3.09, respectively, while they are 0.96, 0.63, and 2.95 in the GB titration. Extending the explicit-solvent simulations to 2 ns gives results that are very similar.

We first examine Asp95, for which the explicit-solvent titration was able to reduce the overestimation of pK_a from the GB-based titration from 1.21 to 0.55 units. The major reason for the improvement is related to the strength of the interaction with Lys70. In the crystal structure obtained at pH of 8, the minimum distance between the charge centers on Asp95 and Lys70 is 4.7 Å, which suggests a salt-bridge interaction. Figure 8 gives the probability distribution of the minimum distance between the charge centers from the explicit-solvent and the GB simulations. Although the average distance is identical at 6.1 Å, the difference lies in the distribution. The GB simulation sampled a unimodal distribution centered around 7 Å. By contrast, the explicit-solvent simulation sampled two distinct populations, one centered at 2.8 Å, representing the conformations where Asp95 and Lys70 are closely associated, and another one centered at 7.1 Å, representing the conformations where the two side chains are rotated away from each other. The bimodal distribution seen in the explicit-solvent simulation is a direct result of including discrete solvent molecules and reflects a more realistic description of the ion pair interaction. However, the GB simulation neglects solvent granularity and models the ion pair interaction in a mean-field manner, which results in a less tight salt-bridge pairing and an underestimation of the pK_a shift for Asp95.

Another case where the inclusion of explicit solvent resulted in the more accurate pK_a calculation is for Asp77. The experimental measurement provides an upper bound of 2.2 for the pK_a. In the explicit-solvent simulation, the pK_a was calculated to be in the correct range, but in the GB simulation, the pK_a shift was underestimated by at least 1 pH unit. Asp77 is within a hydrogen-bond distance of two backbone amide hydrogens of Asn119 and Thr120, which are located in a loop connecting a β-sheet motif to an α-helix (Figure 9, upper left snapshot). In Figure 9 we monitor the minimum distance between the carboxylate oxygens of Asp77 and the backbone amide hydrogen of Asn119 or Thr120. In the explicit-solvent simulation the distance was stable,

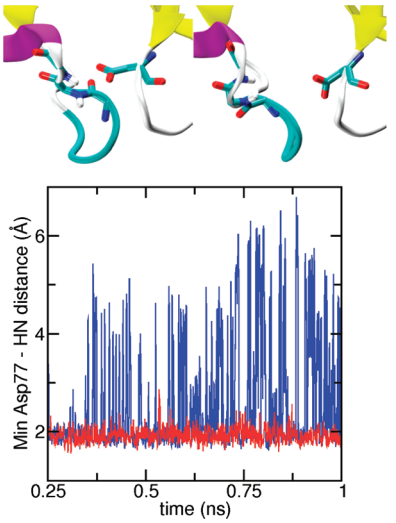

Figure 9. Comparison of the local environment of Asp77 of SNase from the explicit- and implicit-solvent titrations at pH 3. Upper panel: In the initial structure Asp77 forms backbone hydrogen bonds with Asn119 and Thr120 (left snapshot). These interactions were broken in the GB simulation (right snapshot). Lower panel: Time series of the minimum distance between the carboxylate oxygens of Asp77 and the backbone amide hydrogen of Thr119 or Asn120 from the explicit-solvent (red) and GB (blue) simulations at pH of 3.

fluctuating around 2 Å during the entire trajectory, revealing that the backbone hydrogen bonding between Asp77 and Asn119/Thr120 was intact. However, in the GB simulation, this interaction was disrupted as a result of the high mobility of the aforementioned loop (see Figure 9, upper right snapshot). This analysis suggests that the underestimation of the pK_a shift for Asp77 in the GB simulation is due to the distortion of local structure.

The largest pK_a error from the explicit- and implicit-solvent titrations is for Asp21, which interacts with Asp19 on the other end of the β-hairpin. NMR titration data showed two distinct transitions for the two residues.[42] The major transitions have the pK_a of 2.21, assigned to Asp19, and 6.54, assigned to Asp21.[42] The latter is the only upward shifted pK_a relative to the model value for SNase. Both the explicit- and implicit-solvent titrations were not able to reproduce the direction of the pK_a shift for Asp21 and underestimated the pK_a by about 3 pH units, although the explicit-solvent simulation was able to differentiate between the two pK_a's. During the explicit-solvent simulation at pH of 3, the average distance between the carboxylate oxygens of both residues was 3.7 Å. This close proximity was stabilized by a persistent hydrogen bond between the carboxylate oxygen of Asp19 and the backbone amide nitrogen of Asp21. However, the coupled titration behavior with two transitions was not observed when fitting the data for either Asp19 or Asp21. The only indication of coupling was a low Hill coefficient (0.56) for Asp19, which indicates anticooperativity, consistent with experiment.[42] We also examined the GB titration data. The interaction between Asp19 and Asp21 was very strong but both residues titrated with the same pK_a, and the Hill coefficients were about 1. Thus, compared to the GB titration, the explicit-solvent simulation was able to provide, to some extent, the description of the coupled proton binding events for Asp19 and Asp21. However, the explicit-

Table 5. Calculated and Experimental pK_a Values in HEWL

residue	explicit solvent[b]	GB	expt[a]
Glu7	2.67 ± 0.01 (2.69)	2.58 ± 0.06	2.6 ± 0.2
His15	6.64 ± 0.10 (6.60)	5.34 ± 0.47	5.5 ± 0.2
Asp18	3.05 ± 0.13 (3.15)	2.94 ± 0.01	2.8 ± 0.3
Glu35	7.19 ± 0.15 (6.83)	4.35 ± 0.18	6.1 ± 0.4
Asp48	1.57 ± 0.48 (1.77)	2.84 ± 0.15	1.4 ± 0.2
Asp52	2.88 ± 0.08 (3.21)	4.56 ± 0.02	3.6 ± 0.3
Asp66	1.47 ± 0.60 (0.46)	1.15 ± 0.43	1.2 ± 0.2
Asp87	1.48 ± 0.41 (1.46)	2.03 ± 0.07	2.2 ± 0.1
Asp101	2.99 ± 0.09 (3.06)	3.27 ± 0.32	4.5 ± 0.1
Asp119	2.85 ± 0.05 (2.98)	2.45 ± 0.13	3.5 ± 0.3
CT-Leu	1.95 ± 0.37 (1.89)	2.20 ± 0.14	2.7 ± 0.2
avg abs dev	0.70 (0.70)	0.72	
rmsd	0.84 (0.80)	0.93	
max abs dev	1.50 (1.44)	1.75	

[a] Consensus pK_a's based on NMR titration using multiple nuclei.[43] [b] Values in parentheses were obtained from the 2 ns simulation.

solvent simulation was not able to fully capture the negative cooperativity, which may be due to insufficient sampling.

HEWL. The last protein we consider is HEWL, which has been used as a standard test system for many pK_a prediction methods.[52,53] Also, the most recent study of Nielsen and coworkers, where a consensus set of pK_a's was derived from pH-dependent chemical shifts of different nuclei, makes HEWL the most vetted protein pK_a benchmark system available.[43] Table 5 lists the calculated pK_a's from the explicit- and implicit-solvent titrations. Overall, the calculated pK_a's from the explicit-solvent titration are closer to experiment than the GB titration. The rms as well as the average absolute and maximum deviations in the explicit-solvent titration are 0.84, 0.70, and 1.50, respectively, while they are 0.93, 0.72, and 1.75, respectively, in the GB titration. Below we examine the cause for the significant differences between the explicit- and implicit-solvent titration data for residues Glu35 and Asp52.

The catalytic residues of HEWL are Glu35 and Asp52 reside at the interface between two domains and have the concensus pK_a's of 6.1 and 3.6, respectively. The experimental range of pK_a's calculated from chemical shifts of different nuclei was 6.0−6.8 for Glu35 and 3.4−4.0 for Asp52.[43] The pK_a's from the explicit-solvent simulation are 7.19 and 2.88, while those from the GB simulation are 4.35 and 4.56, respectively. Thus, considering the model values of 4.4 and 4.0 for Glu and Asp, the calculated pK_a shifts are in the correct direction in the explicit-solvent simulation but wrong in the GB simulation. Since the optimum pH for the activity of HEWL is around 5,[54] the pK_a calculation using the explicit-solvent CPHMD method is able to offer the correct protonation or charge states for the catalytic residues, which is not the case with the GB-based method. We note that the previous GB-based CPHMD simulations with the temperature-based replica-exchange protocol gave a correct direction of the pK_a shift for Glu35.[26] We examined the trajectory to delineate the cause for the significantly different pK_a's. In the GB simulation, there is a significant rearrangement of the native structure. We plot the radius of gyration vs the heavy atom rmsd using the explicit- and implicit-solvent simulation data (Figure 10). The conformations in explicit solvent have rmsd

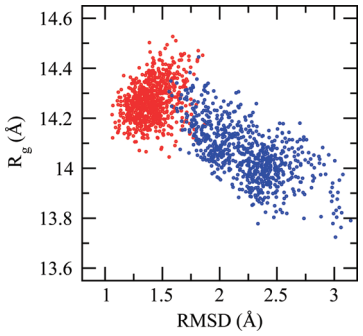

Figure 10. Comparison of the conformational states of HEWL sampled in the explicit-solvent (red) and GB (blue) simulations at pH 6.

values, with respect to the crystal structure, ranging from 1.1 and 1.6 Å, and R_g values ranging from 14.1 to 14.4 Å. However, the conformations in the GB simulation have much larger rmsd (1.6−2.8 Å) and much smaller R_g (13.8−14.2 Å), which suggests a significant compaction and global deviation from the crystal structure. This global rearrangement of structure is propagated to the local conformational environment around the active-site residues, which can be seen from the differences in the solvent exposure of side chains. At pH 6, Glu35 has a SASA of 18.9 Å^2 in the explicit-solvent simulation, which is similar to the value of 10 Å^2 based on the crystal structure but much smaller than the value of 38.9 Å^2 from the GB simulation. The significant increase in solvent exposure for Glu35 in the GB simulation leads to an overestimation of the self-solvation energy of Glu35 and thus an underestimation of the upward pK_a shift. For Asp52 the story is exactly reversed. The solvent exposure of Asp52 is underestimated in the GB as compared to the explicit-solvent simulation. At pH 4, the average SASA of Asp52 is 2.4 Å^2 in the GB simulation, whereas it is 25.4 Å^2 in the explicit-solvent simulation, which is much closer to the initial value of 26.6 Å^2. Therefore, the self-solvation energy of Asp52 is underestimated in the GB simulation leading to a calculated pK_a value that is too high. Thus, HEWL is a clear case where elimination of the "conformational" error introduced by GB can dramatically improve the accuracy of pK_a calculations for residues of biological significance.

DISCUSSION

Like temperature and pressure, solution pH is another important experimental condition that needs to be taken into account in molecular simulations in order to accurately capture physical reality. Motivated by the recent success of the GB implicit-solvent based CPHMD method in the accurate pK_a predictions and mechanistic studies of pH-dependent conformational dynamics of proteins, we have developed a robust approach to extend the CPHMD framework to explicit-solvent molecular dynamics simulations. In this approach, the explicit-solvent force field is used to drive conformational dynamics, while the GB model is used to efficiently estimate the role of solvent in modulating the cost of electrostatic free energy for protonation/deprotonation. The resulting explicit-solvent CPHMD method offers an increased accuracy and a wider applicability as compared to the GB-based CPHMD method while retaining the efficiency and the robustness of the capability for proton titration. To overcome a critical hurdle related to the slow convergence of pK_a calculations, which has plagued

CPHMD and other constant pH methodologies, we have implemented a replica-exchange protocol based on the pH-biasing energy to directly accelerate protonation-state sampling. Remarkably, due to the tight coupling between titration and conformational degrees of freedom, this protocol also led to significant enhancement in conformational sampling, allowing pK_a's to converge within 1 ns for small model compounds and large proteins. The random errors in the calculated pK_a's for model compounds were about or below 0.1 pH units.

To benchmark the accuracy of the explicit-solvent based CPHMD method, we have calculated pK_a's for five proteins and compared with results from the GB-based method and experiment. We found that the explicit-solvent titrations resulted in an average absolute error of 0.53 and rmsd of 0.74, on par with those from the GB-based titrations. However, by bringing the outliers closer to experimental values, the explicit-solvent method offers significantly improved correlation with experiment as compared to the GB-based method. Detailed analysis revealed that this improvement is due to more accurate conformational sampling in explicit solvent. For example, the explicit-solvent simulation preserved the structural integrity of the loop region, bringing the calculated pK_a of Asp77 from SNase closer to experiment. Compaction of HEWL in the implicit-solvent simulation caused distortion of the active site and large deviations in the calculated pK_a values for Glu35 and Asp52, while explicit-solvent simulation preserved the native conformation leading to a correct prediction of the protonation states at the optimum pH value for catalytic activity. Including solvent granularity enabled a more realistic description of ion-pair interactions, as was the case for Asp95 of SNase, where the explicit-solvent simulation gave a bimodal distribution representing both the close-range and solvent-separated interactions with Lys70, which resulted in a more accurate estimate of pK_a. Finally, in the explicit-solvent simulation the hydrophobic cluster in BBL showed an increased mobility relative to the GB simulation, allowing His166 to be partially exposed to solvent, which resulted in a reduction in the pK_a shift due to desolvation penalty. The latter aspect is somewhat surprising, but is compatible with previous GB simulation studies revealing overly rigid hydrophobic assemblies.[23,24] It is also consistent with the experimental evidence[55] and previous simulation study[56] suggesting water penetration into the hydrophobic core of SNase. Although in the presented cases, the differences between the explicit-solvent and GB-based pK_a results are small (all within 0.5 pH units), our unpublished data shows that the explicit-solvent method offers improvement as high as 4 pH units for the worst prediction cases in the engineered mutants of SNase (Wallace and Shen, unpublished data).

While the results demonstrated in this work are encouraging, we note that several potential issues merit attention. First, a potential delay in the response of solvent reorganization to protonation/deprotonation may lead to unfavorable interactions or inaccuracy in the solvation energetics of the titration site. This problem can be effectively avoided by allowing a few additional dynamics steps between titration updates to allow relaxation of solvent around the titrating site, as has been demonstrated in the discrete constant pH techniques.[11] Also, we identified a small bias toward the charged form in the titration of Asp, Glu and His residues due to the occasionally prolonged residence time of the two degenerate protonation states (doubly deprotonated in the case of Asp or Glu and doubly protonated in the case of His). Although the effect of this systematic error on the calculated pK_a shifts is likely minimal, it is clearly a limitation that needs to addressed in the future. Finally, the accuracy of pK_a calculations is still limited by the accuracy of the GB model to determine the deprotonation free energy. The largest deviation and the single outlier found in this work is Asp21 in SNase, where both explicit- and implicit-solvent simulations were not able to reproduce the direction of the positive pK_a shift, and underestimated the pK_a by 3 pH units. NMR data showed that the titration of Asp21 is coupled to that of Asp19, which has a negative pK_a shift. Although the explicit-solvent simulation was able to differentiate between the two pK_a's, it could not quantitatively reproduce the extent of the negative cooperativity in proton binding. One possible cause is that more exhaustive sampling may be required to fully capture coupled titration events. This issue deserves further investigation in our future studies. Another aspect that deserves further investigation is related to the effect due to ions. In the current work and previous GB-based CPHMD studies, an approximated Debye−Hückel model is applied in the GB electrostatic calculation to account for the bulk effect of salt screening, which may not be accurate for highly charged systems such as nucleic acids where local charge density can be very high. Finally, in order to apply the explicit-solvent CPHMD to studies of large-scale conformational changes, it may become necessary to combine with a method for global enhancement of conformational sampling such as the temperature-based replica-exchange scheme. Despite these remaining limitations, the current accuracy and precision of the explicit-solvent based CPHMD technique are encouraging, considering the fact that experimentally determined pK_a's can deviate by 0.5−1 pH units depending on the nuclei monitored.[43] Thus, we anticipate that explicit-solvent CPHMD simulations will emerge as a practical tool for gaining novel insights into protonation-related phenomena that are ubiquitous in biology and chemistry. Examples include the mechanism of proton channels, drug-efflux pumps, pH-dependent catalytic reactions of ribozymes, as well as titration behavior of mixed micelle systems.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jana.k.shen@ou.edu. Telephone: (405) 325-0458.

ACKNOWLEDGMENT

Financial support provided by University of Oklahoma and the American Chemical Society Petroleum Research Fund.

REFERENCES

(1) Warshel, A. *Biochemistry* **1981**, *20*, 3167–3177.
(2) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159.
(3) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284–290.
(4) Nielsen, J. E.; Mccammon, J. A. *Protein Sci.* **2003**, *12*, 1894–1901.
(5) Shen, J. K. *J. Am. Chem. Soc.* **2010**, *132*, 7258–7259.
(6) Börjesson, U.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.
(7) Bürgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469–480.
(8) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.
(9) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *Proteins* **2004**, *56*, 738–752.

(10) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.
(11) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.
(12) Wallace, J. A.; Shen, J. K. *Methods Enzymol.* **2009**, *466*, 455–475.
(13) Khandogin, J.; Brooks, C. L., III. *Biophys. J.* **2005**, *89*, 141–157.
(14) Kong, X.; Brooks, C. L., III. *J. Chem. Phys.* **1996**, *105*, 2414–2423.
(15) Khandogin, J.; Chen, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 18546–18550.
(16) Khandogin, J.; Brooks, C. L., III. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 16880–16885.
(17) Khandogin, J.; Raleigh, D. P.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.
(18) Chen, J.; Im, W.; Brooks, C. L., III. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.
(19) Okur, A.; Wickstrom, L.; Simmerling, C. *J. Chem. Theory Comput.* **2008**, *4*, 488–498.
(20) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 115–127.
(21) Chen, J.; Brooks, C. L., III. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.
(22) Wallace, J. A.; Shen, J. K. *Biochemistry* **2010**, *49*, 5290–5298.
(23) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702–4709.
(24) Shen, J. K. *Biophys. J.* **2010**, *99*, 924–932.
(25) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.
(26) Khandogin, J.; Brooks, C. L., III. *Biochemistry* **2006**, *45*, 9363–9373.
(27) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
(28) Im, W.; Lee, M. S.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
(29) Srinivasan, J.; Trevathan, M. W.; Beroza, P.; Case, D. A. *Theor. Chem. Acc.* **1999**, *101*, 426–434.
(30) Börjesson, U.; Hünenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.
(31) Nadler, W.; Hansmann, U. H. E. *Phys. Rev. E* **2007**, *75*, 026109.
(32) Meng, Y.; Roitberg, A. E. *J. Chem. Theory Comput.* **2010**, *6*, 1401–1412.
(33) Nadler, W.; Hansmann, U. H. E. *J. Phys. Chem. B* **2008**, *112*, 10386–10387.
(34) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.
(35) Okamoto, Y. *J. Mol. Graphics Modell.* **2004**, *22*, 425–439.
(36) Nozaki, Y.; Tanford, C. *Methods Enzymol.* **1967**, *11*, 715–734.
(37) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045–8056.
(38) Bi, Y. *Studies of the folding and stability of the villin headpiece subdomain*, Ph.D. Thesis, Stony Brook University: Stony Brook, NY, 2008.
(39) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(40) Arbely, E.; Rutherford, T. J.; Sharpe, T. D.; Ferguson, N.; Fersht, A. R. *J. Mol. Biol.* **2009**, *387*, 986–992.
(41) Kuhlman, B.; Luisi, D. L.; Young, P.; Raleigh, D. P. *Biochemistry* **1999**, *38*, 4896–4903.
(42) Castañeda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; García-Moreno, E., B. *Proteins* **2009**, *77*, 570–588.
(43) Webb, H.; Tynan-Connolly, B. M.; Lee, G. M.; Farrell, D.; O'Meara, F.; Søndergaard, C. R.; Teilum, K.; Hewage, C.; McIntosh, L. P.; Nielsen, J. E. *Proteins* **2011**, *79*, 685–702.
(44) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graphics Modell.* **2004**, *22*, 377–395.
(45) Mackerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400–1415.
(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
(47) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.
(48) Bogusz, S.; Cheatham, T. E., III.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070–7084.
(49) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
(50) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.
(51) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.
(52) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.
(53) Khandogin, J. Modeling protonation equilibria in biological macromolecules. In *Multi-scale quantum models for biocatalysis*York, D. M., Lee, T.-S., Eds.; Springer: New York, 2009; Chapter 10, pages 261−284.
(54) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180–1184.
(55) Denisov, V. P.; Schlessman, J. L.; García-Moreno, E., B.; Halle, B. *Biophys. J.* **2004**, *87*, 3982–3994.
(56) Damjanović, A.; García-Moreno, B.; Lattman, E. E.; García, A. E. *Proteins* **2005**, *60*, 433–449.
(57) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2006**, *15*, 1214–1218.