

# Measuring the successes and deficiencies of constant pH molecular dynamics: A blind prediction study

Sarah L. Williams,<sup>1\*</sup> Patrick G. Blachly,<sup>1</sup> and J. Andrew McCammon<sup>1,2,3,4</sup>

<sup>1</sup>Department of Chemistry & Biochemistry, University of California San Diego, La Jolla, California 92093-0365

<sup>2</sup>Center for Theoretical Biological Physics, University of California San Diego, La Jolla, California 92093

<sup>3</sup>Howard Hughes Medical Institute, University of California San Diego, La Jolla, California 92093-0365

<sup>4</sup>Department of Pharmacology, University of California San Diego, La Jolla, California 92093-0365

## ABSTRACT

A constant pH molecular dynamics method has been used in the blind prediction of  $pK_a$  values of titratable residues in wild type and mutated structures of the Staphylococcal nuclease (SNase) protein. The predicted values have been subsequently compared to experimental values provided by the laboratory of García-Moreno. CpHMD performs well in predicting the  $pK_a$  of solvent-exposed residues. For residues in the protein interior, the CpHMD method encounters some difficulties in reaching convergence and predicting the  $pK_a$  values for residues having strong interactions with neighboring residues. These results show the need to accurately and sufficiently sample conformational space in order to obtain  $pK_a$  values consistent with experimental results.

Proteins 2011; 79:3381–3388.  
© 2011 Wiley-Liss, Inc.

**Key words:** constant pH molecular dynamics (CpHMD);  $pK_a$  prediction; implicit solvation; Monte Carlo.

## INTRODUCTION

It is well established that the structure and function of a protein are highly dependent on the pH of its surrounding environment. The  $pK_a$  of a titratable residue, which is heavily influenced by interactions with neighboring residues within a protein, governs the protonation state of that residue for a given solution pH. Changes in protonation state within a protein manifest as alterations to the charge distribution of the titratable residue, influencing the electrostatics of the protein environment. Protonation equilibria are thus closely linked with protein conformation, evidence of which is the sensitivity of proteins to denaturation at extreme pH.

The interplay between protonation state and protein conformation is not accounted for in conventional molecular dynamics (MD) simulations. Currently, these simulations employ fixed, predetermined protonation states for titratable residues, which are generally chosen according to the  $pK_a$  value of the respective residue when isolated in solution. This method of protonation state assignment can be a severe approximation, as the  $pK_a$  values of titratable residues are frequently shifted from that of the isolated residue in solution. Furthermore, protonation states are not constant, but rather exist in equilibria, subject to the changing electrostatic environment surrounding the titratable group. Therefore, incorporating pH as an input variable in MD simulations is highly desirable, as it would allow a more accurate study of pH-coupled protein dynamics, such as ligand binding and protein folding.

Over the past few decades, a number of theoretical methods have been developed to try to accurately determine the protonation states of titratable residues in proteins. One class of methods utilizes static protein structures and employs a Poisson–Boltzmann approach for the calculation of electrostatics.<sup>1–3</sup> However, the use of static structures is thought to be a major contributor to discrepancies observed in the calculation of  $pK_a$  shifts, as the conformational changes in the protein induced by change in residue protonation state are not taken into account. More recently, these methods have been improved by including descriptions of conformational variability, with adaptations to account for dielectric heterogeneity<sup>4,5</sup> and inclusion of conformational flexibility.<sup>6–9</sup> Notably, Warshel and coworkers were the first to employ MD methods to improve calculation of  $pK_a$  values in proteins, with their electrostatic protein dipoles Langevin dipoles (PDL) model.<sup>10</sup> Other groups have incorporated MD and QM/MM methods

The authors state no conflict of interest.

Sarah L. Williams and Patrick G. Blachly contributed equally to this work.

Grant sponsors: NSF, NIH, HHMI, CTBP, NBCR, NSF supercomputer centers.

\*Correspondence to: Sarah L. Williams, Department of Chemistry & Biochemistry, University of California, San Diego, 9500 Gilman Drive, Mail code 0365, La Jolla, CA 92093-0365. E-mail: swilliam@mccammon.ucsd.edu

Received 10 January 2011; Revised 19 May 2011; Accepted 25 May 2011

Published online 22 July 2011 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.23136

coupled with free energy perturbation techniques for  $pK_a$  calculations.<sup>11–13</sup> A drawback of these techniques is their high reliance on the resolution of the input structure, which renders these methods incapable of calculating  $pK_a$  shifts where protonation is accompanied by large conformational change.

Another class of methods incorporates the important coupling of conformation and protonation state through the use of computational simulations that employ pH as an external thermodynamic parameter.<sup>14–26</sup> These methods are often described as either continuous or discrete constant pH methods, contingent on how titratable protons are considered within the simulation. The former treats protonation state as a continuous titration parameter that advances simultaneously with the atomic coordinates of the system.<sup>19–21</sup> Originally, the implementation of this method used a mean-field approximation, and protonation sites could exist as fractionally occupied. More recently, Lee *et al.*, have developed methods to overcome issues with fractional protonation states, using  $\lambda$ -dynamics with an artificial titration barrier to discourage fractional protonation.<sup>22</sup> Extensions to the work of Lee *et al.*, have incorporated proton tautomerism<sup>23</sup> and enhanced sampling methods to improve convergence.<sup>24</sup>

Discrete constant pH methods avoid non-physical intermediate charge states. These methods use MD simulations for conformational sampling, while sampling different discrete protonation states with periodic Monte Carlo (MC) steps interspersed throughout the MD trajectory.<sup>14–18</sup> The methods employed in this article utilize the constant pH MD (CpHMD) method, originally developed by Mongan *et al.*,<sup>14</sup> which uses generalized Born (GB) implicit solvent. Differences among these methods arise from choice of solvation model and protocols for updating protonation states within the simulation. Although these methods have achieved good results for small protein systems, they can be computationally expensive, and long convergence times have been reported for systems with multiple titration sites. In an attempt to overcome these issues of convergence, use of enhanced sampling methods coupled with constant pH MD, such as constant pH accelerated MD (CpHaMD)<sup>25</sup> and constant pH replica-exchange MD (REX-CpHMD)<sup>26</sup> have been investigated. Results from simulations employing these methods indicate the increased sampling provides improvement over the conventional method.

The previous paragraphs provide only a brief summary of the computational methods available for  $pK_a$  prediction. Further details of these and other methods can be found in the literature, and several reviews have been published.<sup>27–29</sup> In this study, the successes and deficiencies of the CpHMD method have been investigated in the blind prediction of  $pK_a$  values of titratable residues of the WT and mutant forms of the Staphylococcal nuclease (SNase) enzyme based upon comparison to experimental results released after submission to the  $pK_a$  cooperative<sup>30</sup> by Gar-

cía-Moreno and coworkers.<sup>31–37</sup> Particular attention is paid to the differences in electrostatics and, consequently, acid/base properties of exterior and interior residues.

## THEORY

### Constant pH molecular dynamics—Theory background

CpHMD employs MD with GB implicit solvent.<sup>14</sup> Within the simulation, the MD simulation is periodically halted, and a MC step is taken, randomly considering a titratable residue for change in protonation. The transition energy corresponding to this MC step is evaluated according to Eq. (1), which calculates  $pK_a$  with respect to a reference

$$\Delta G = k_B T (\text{pH} - pK_{a,\text{ref}}) \ln 10 + \Delta G_{\text{elec}} - \Delta G_{\text{elec,ref}} \quad (1)$$

compound for the residue of interest. Reference compounds are the isolated titratable residues solvated in water (reference  $pK_a$  values are 3.8 for ASP, 4.3 for GLU, 6.8 for HIS, 9.6 for TYR, and 10.5 for LYS).<sup>14,38,39</sup> In Eq. (1),  $k_B$  is the Boltzmann constant,  $T$  is the temperature, pH is the specified solvent pH,  $pK_{a,\text{ref}}$  is the  $pK_a$  of the reference compound,  $\Delta G_{\text{elec}}$  is the electrostatic energy change for protonation state change of the titratable residue, and  $\Delta G_{\text{elec,ref}}$  is the corresponding electrostatic transition energy for the reference compound. The same GB electrostatics employed in the MD is used for calculating this transition energy, with acceptance of the change in protonation determined by the Metropolis criterion. If the MC move is accepted, the protonation state of the residue will change to the new state, and MD is continued. If not, the simulation will continue with the residue remaining in the unchanged protonation state. CpHMD has been successfully applied in the  $pK_a$  prediction of titratable residues in the Hen Egg White Lysozyme (HEWL) enzyme.<sup>14</sup>

### Titration curve construction and $pK_a$ calculation

The predicted  $pK_a$  values are calculated from performing CpHMD simulations over a range of solution pH values. Assuming the system is ergodic, we assume fractional protonation is given by the amount of time a particular titratable residue spends in its protonated state.<sup>15</sup> Thus the fraction of deprotonated species,  $s$ , for a residue at a specific pH value can be used to predict the  $pK_a$  from a Hill plot [Eq. (2)].<sup>16,20,22,40</sup> Fits to this curve

$$s = \frac{1}{1 + 10^{n(\text{pH} - pK_a)}} \quad (2)$$

allow for estimation of both the  $pK_a$  value as a midpoint of titration, as well as the Hill coefficient,  $n$ , which

describes the cooperativity of various sites with respect to titration.<sup>41</sup> Illustrated in Reference 40, for example, the usefulness of the Hill equation resides in its ability to provide a good prediction of the midpoint  $pK_a$  value, even when the fit is inaccurate at the tails of the titration curve.<sup>40</sup>

## METHODS

### Test system: Staphylococcal nuclease

Staphylococcal nuclease (SNase) is a highly charged protein, which has generated difficulty in obtaining accurate structure-based  $pK_a$  predictions.<sup>42</sup> The structures of the wild-type and mutant proteins of the SNase system are provided for this study by the lab of Garcia-Moreno *et al.*, who measured the  $pK_a$  shifts of the titratable residues using NMR spectroscopy.<sup>31–37</sup> Along with other computational groups, we have computed blind  $pK_a$  predictions for residues of wild-type SNase (PDB ID: 1STN or 1SNC), the SNase mutant  $\Delta$ +PHS (PDB ID: 3bdc), and various mutants from the  $\Delta$ +PHS parent protein (referred to as calculated results in this study).  $\Delta$ +PHS is unique in that it is a hyperstable, acid-resistant SNase mutant with five substitutions (G50E, V51N, P117G, H124L, and S128A) and a deletion of residues 44–49.<sup>31,32</sup> Garcia-Moreno directed this effort, holding experimental  $pK_a$  determinations from those making predictions and picking residues of interest for  $pK_a$  prediction (hereafter referred to as experimental results in this study). In total, approximately 93 structures of the wild-type (WT) and mutant SNase have been provided.<sup>30</sup> Owing to time constraints, however, our CpHMD  $pK_a$  predictions have not been carried out on the entire set of provided structures, the subset of which were studied and submitted as blind predictions shown in Table I.

### CpHMD simulations

The standard CpHMD method has been implemented in AMBER10 molecular dynamics program. All simulations are conducted with the AMBER99SB force field<sup>43</sup> and the GB solvent model  $igb=2$ ,<sup>44–46</sup> using a 30 Å cut-off value for nonbonded interactions and computation of effective Born radii calculations. Similar to experimental conditions, salt concentrations are set to either 0.1 or 1.0M. The SHAKE algorithm constrains all bonds involving hydrogen with a time step of 2 fs,<sup>47</sup> and temperature is maintained at 300 K using the Berendsen temperature coupling method with a time constant of 2 ps.<sup>48</sup> A period of 10 fs of MD separates the MC trials. With these parameters, a 10 ns CpHMD simulation takes approximately 72 h using 16 Xeon X5650 2.67GHz processors.

All simulations begin from the crystal structure coordinates of the WT and  $\Delta$ +PHS SNase systems provided by Garcia-Moreno *et al.*, from which specific titratable

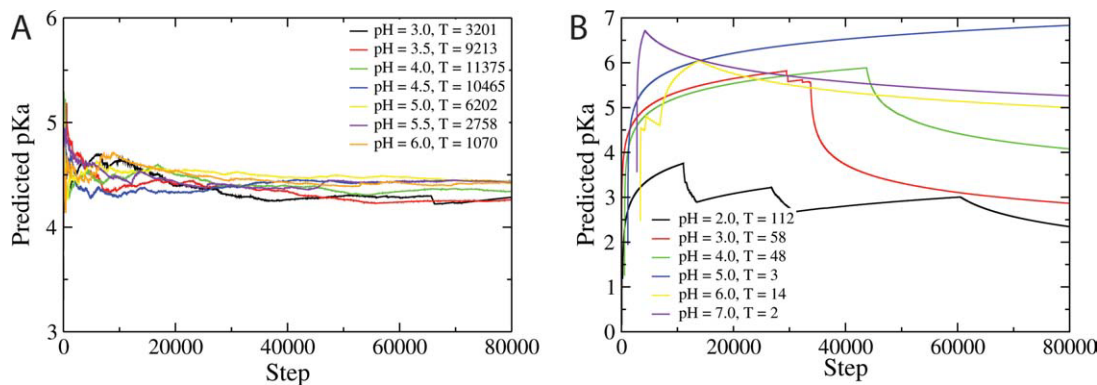
**Table I**

Predicted and Experimental Values for Various Residues from the WT SNase,  $\Delta$ +PHS, and  $\Delta$ +PHS Mutant Proteins<sup>30,31</sup>

Protein	Residue	Experimental $pK_a$	Predicted $pK_a$	(Pred.–Exp) $pK_a$ offset	(Pred.–Model) $pK_a$ offset
WT	HIS8	6.52	5.67 ± 0.04	−0.85	−1.1
	HIS46	5.86	6.8 ± 0.3	0.7	0.0
	HIS121	5.30	7.0 ± 0.1	1.7	0.2
	HIS124	5.73	6.0 ± 0.1	0.3	−0.8
$\Delta$ +PHS	ASP19	2.21	4.1 ± 1.1	0.9	0.3
	ASP21	6.54	–	–	–
	ASP40	3.87	3.1 ± 0.1	−0.8	0.7
	ASP77	<2.2	3.6 ± 0.2	>1.1	−0.2
	ASP83	<2.2	2 ± 8	–	−2
	ASP95	2.16	3.6 ± 0.1	1.4	−0.2
	GLU10	2.82	4.4 ± 0.2	1.6	0.1
	GLU43	4.32	1 ± 5	−0.9	−3
	GLU52	3.93	4.3 ± 0.2	0.4	0.0
	GLU57	3.49	4.3 ± 0.1	0.8	0.0
	GLU67	3.76	4.39 ± 0.03	0.6	0.09
	GLU73	3.31	4.2 ± 0.1	0.9	−0.1
	GLU75	3.26	4.0 ± 0.1	0.7	−0.3
	GLU101	3.81	3.5 ± 0.2	−0.3	−0.8
	GLU122	3.89	3.8 ± 0.1	−0.1	−0.5
GLU129	3.75	4.28 ± 0.04	0.6	−0.02	
GLU135	3.76	4.2 ± 0.1	0.4	−0.1	
F34E	GLU34	7.30	5.9 ± 0.1	−1.4	1.6
F34K	LYS34	7.10	2 ± 5	−5	−3
G20D	ASP20	<4.0	2 ± 2	−2	−2
G20E	GLU20	<4.5	4.1 ± 0.3	–	−0.2
G20K	LYS20	>10.4	8.6 ± 0.2	<−1.8	−1.8
L25D	ASP25	6.80	4.8 ± 0.3	−2.0	1.0
L36D	ASP36	7.90	5 ± 3	−3	1
L37D	ASP37	<4.0	–	–	–
V23D	ASP23	6.8	3 ± 2	−4	−1
V23E	GLU23	7.1	6.4 ± 0.1	−0.7	2.1
V23K	LYS23	7.40	7.3 ± 0.6	−0.1	−3.1

The difference between experimental and model compound  $pK_a$  values for  $\Delta$ +PHS mutants. Model compound  $pK_a$  values: 3.8 (ASP), 4.3 (GLU), 6.8 (HIS), and 10.4 (LYS).<sup>14,38,39</sup>

residues were chosen for the blind prediction study (Table I). For the blind predictions performed on SNase systems where a single ASP or GLU residue has been highlighted as the residue of interest, CpHMD simulations of 10 ns in length have been performed in the solution pH range 2.0–7.0 at 0.5 pH unit intervals, titrating only acidic residues. For these simulations, HIS residues are allowed to titrate from pH 4.5 to pH 7.0. In systems where a LYS or TYR is highlighted as the residue of interest, simulations have been carried out in the pH range 7–10.5, where HIS, LYS, and TYR residues are set to titrate. The exclusion of HIS residues from the most acidic simulations is justified, as the  $pK_a$  of the HIS reference is around 6–7.<sup>39</sup> In most cases, it is safe to assume all ASP and GLU residues are deprotonated above pH 7 and all LYS and TYR are protonated below pH 7, allowing exclusion of these residues from titration in these respective pH regions. Models for the terminal residues have not yet been developed for this system, so these res-



**Figure 1**

Plots of predicted  $pK_a$  over the duration of CpHMD simulations for (a)  $\Delta$ +PHS GLU52 (experimental  $pK_a = 3.93$ )<sup>31</sup> and (b)  $\Delta$ +PHS L36D (experimental  $pK_a = 7.90$ ).<sup>35</sup> The number of protonation state transitions (T) are given in the figure legend for each system.

idues are set to their most likely protonation states at neutral pH, with the N-terminus protonated and the C-terminus deprotonated. All non-titrating residues are set to their expected protonation states.

### Simulations conducted after publishing of experimental results

To understand why some predictions fail to reproduce the experimental results, further CpHMD simulations have been conducted, as indicated in the proceeding sections. In particular, the pH range at which simulations were originally performed is extended to account for residues that deviate the most from their reference value. In cases where convergence has been determined to be problematic, extended simulations (>10 ns) do not appear to improve predictions (results not shown).

## RESULTS

### Titration curves

Titration curves are obtained from CpHMD simulations for 32 titratable residues of the WT,  $\Delta$ +PHS, and mutant  $\Delta$ +PHS SNase systems. The experimental  $pK_a$  values for the WT protein were published prior to the predictions, but those for the mutant proteins were withheld until blind predictions were made (Table I). From Eq. (2),  $pK_a$  values are calculated along with the standard errors of regression for curve fits to the Hill equation in Eq. (2) (Table I).<sup>49</sup> In the instances of ASP21 and L37D,  $pK_a$  values cannot be computed due to the lack of transitions between protonated and deprotonated forms.

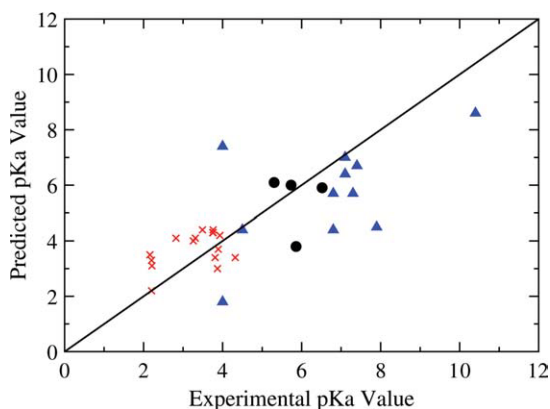
A representative plot of calculated  $pK_a$  over time is given in Figure 1 for both a surface residue for which the CpHMD predicts  $pK_a$  accurately ( $\Delta$ +PHS, GLU52) and the  $\Delta$ +PHS mutant L36D, for which the  $pK_a$  prediction of the interior residue ASP36 deviates by more than three

$pK_a$  units from the experimental result. In the former case, the  $pK_a$  converges rapidly, whereas ASP36 in  $\Delta$ +PHS L36D is indicated to not achieve convergence over the duration of the simulations.

In assessing the convergence of a system, it is interesting to observe the trend in the number of transitions between protonated and deprotonated states as a function of pH. In systems that are well converged (e.g.,  $\Delta$ +PHS, GLU52), the greatest number of transitions between deprotonated and protonated states within the CpHMD scheme are found for the simulation conducted at a pH nearest to the calculated  $pK_a$  value. Simulations conducted at pH values far from the predicted  $pK_a$  encounter fewer transitions between protonation states, as is to be expected from the acceptance criteria defined in Eq. (1). This is not the case for certain systems (e.g.,  $\Delta$ +PHS L36D), where convergence is a problem. Thus the presence of a clear distribution of transitions across the different pH values simulated, peaked at the pH nearest the predicted  $pK_a$ , may be an indicator of how well converged the system is.

### Experimental validation

Following the blind predictions, García-Moreno and coworkers have released experimental results for comparison to predicted  $pK_a$  values (Table I).<sup>30</sup> In summary, CpHMD simulations calculate the  $pK_a$  values of 17 residues to within 1  $pK_a$  unit of the experimental value, 9 residues within 2 units, 1 residue within 3 units, and 2 residues within 4 units. Residues in the  $\Delta$ +PHS protein chosen for analysis are surface residues. All of these residues remain solvent-exposed throughout the CpHMD simulation and are generally well predicted with respect to experiment (Table I). From Figure 2, it is clear that the predictions with the largest deviation from the experimental values are for  $\Delta$ +PHS variants that have residues



**Figure 2**

Plot of predicted versus experimental  $pK_a$  values for WT SNase (●),  $\Delta$ +PHS (x, exterior residues), and  $\Delta$ +PHS mutants ( $\Delta$ , internal residues).<sup>31–37</sup> The line  $y = x$  represents accurate prediction of the experimental  $pK_a$ .

located within the hydrophobic interior of the protein (e.g., L37D, L36D). These residues have also been found experimentally to have the largest shifts in  $pK_a$  from their reference values (Table I). Errors with respect to the experimental  $pK_a$  are shown in Table I, with predictions within ranges of experimental  $pK_a$  considered to have zero error.

As described in Table I, CpHMD simulations have correctly predicted the experimental trends for the majority of these buried residues, three within 1  $pK_a$  unit of the experimental result (G20E, V23E and V23K). The simulations predict the  $pK_a$  values of F34E/K, L25D, L36D, and V23E/K to be shifted from their model values in the direction of favoring the neutral residue at physiological pH; although, for some of these residues, the shift in the predicted  $pK_a$  is not as large as that found experimentally (Table I).

## DISCUSSION

### Residues with $pK_a$ predictions greater than 1 $pK_a$ unit from experimental

Since the release of experimental results, further simulations have been carried out to investigate why our methods predict  $pK_a$  values that deviate more than 1  $pK_a$  unit from experimental results. For this article, we have chosen a selection of residues that illustrate problems with the application of the CpHMD method to these specific systems.

#### $\Delta$ +PHS: ASP21

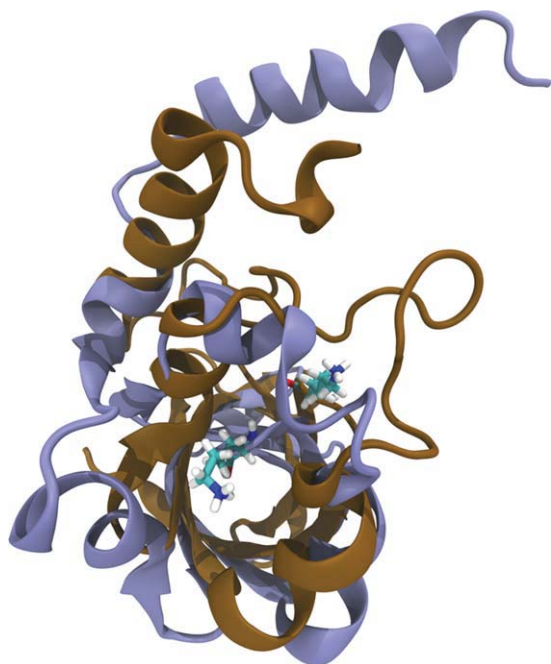
In the  $\Delta$ +PHS mutant, the residue ASP21 is a notable exception to the good performance of CpHMD in pre-

dicting  $pK_a$  values of surface residues. This problem arises from a lack of transitions between protonated and deprotonated states. In this case, longer simulations fail to alleviate the problem, likely due to the existence of a strong, charged hydrogen bond interaction between ASP19 and ASP21 preventing changes in protonation state from occurring. Consistent with our results, García-Moreno and coworkers have needed to apply two-site binding isotherms to properly describe the experimental titration of these interacting residues and have also noted the difficulty in predicting the  $pK_a$  for ASP21 computationally.<sup>31</sup> Similar problems arise in the simulation of L37D, indicating that sampling of protonation states is critical to the performance of the CpHMD method. Use of enhanced sampling techniques to allow the system to sample other protonation states may be necessary for accurate  $pK_a$  predictions in conventional simulations where strong interactions persist.

#### $\Delta$ +PHS G20K

For the mutant  $\Delta$ +PHS G20K, the CpHMD method predicts a  $pK_a$  of 8.6, nearly two  $pK_a$  units lower than the experimental value ( $>10.4$ ).<sup>33</sup> This lysine residue sufficiently sampled protonation space, encountering more than 600 transitions over the duration of each simulation. The trajectories of these simulations incur large motions indicative of protein instability. The root mean square distances (RMSD) with respect to the starting structure for these simulations do not converge at any solution pH, largely influenced by the winding and helical motion of the last 20 residues of C-terminus (Fig. 3). To further probe this conformational change, we performed conventional MD simulations with set protonation states computed by the program PRO  $pK_a$ <sup>50</sup> for pH 7–10 at intervals of 0.5. The conventional MD simulations similarly suffer from protein instability near neutral pH (pH 7 and pH 8), although take longer to encounter it than CpHMD simulations. At higher pH, the terminal helix in the G20K protein does not incur the same motion observed at neutral pH and in CpHMD simulations. These simulations show the sensitivity of the G20K protein toward change in protonation state, and in order to achieve results closer to experiment with the CpHMD method, it may be necessary to spatially constrain the termini. These findings indicate that although increased sampling is desirable and may be achieved in certain systems, it is important that the correct conformational space is sampled to attain an accurate prediction of  $pK_a$ .

In further probing the problems involving predicting the  $pK_a$  for G20K, it is noteworthy that other mutations at site 20 generate similar instabilities (e.g., G20D and G20E). Having an acidic residue at site 20, however, does not affect the  $pK_a$  prediction to the same extent. Visual analysis of trajectories for the G20D protein reveals that hydrogen bonds from Thr-29 persist throughout the sim-



**Figure 3**

Conformational change encountered by the  $\Delta$ +PHS G20K protein at the start (copper) and end (purple) of CpHMD simulation performed at pH 8.5.

ulations, likely lowering its  $pK_a$  (Table I). Similarly, G20E forms transient hydrogen bonds with Thr-29. All mutated residues at site 20 sample conformational space that is solvent-exposed, in addition to time spent buried in the protein interior. While this explains the propensities for G20D and G20E to exist in their charged states, it fails to explain the shift in  $pK_a$  for G20K.

#### $\Delta$ +PHS F34E

CpHMD simulations performed on the  $\Delta$ +PHS F34E mutant consistently obtain a predicted  $pK_a$  (5.7) lower than experiment (7.30).<sup>32</sup> The stability of this particular mutant shows great sensitivity to the pH of the simulation, with large conformational changes occurring at acidic pH (Fig. 4). Nevertheless, the  $pK_a$  values calculated at neutral pH—closer to the  $pK_a$  of the residue—still underestimate the experimental  $pK_a$  despite undergoing a large number of transitions between protonated and deprotonated states. Upon visualization of this structure, it is notable that the carboxylate of GLU34 forms salt bridges with an adjacent arginine residue (ARG81), which causes this residue to favor its deprotonated state. The simulations may not sample enough conformational space owing to the persistence of this salt bridge, therefore leading to a predicted  $pK_a$  value lower than the experimental result. Enhanced sampling techniques may provide the means to allow the system to escape this

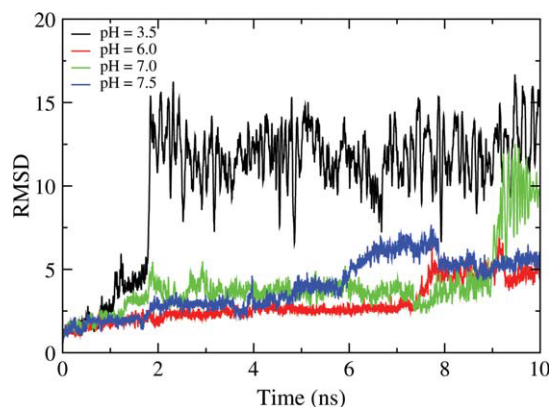
GLU34-ARG81 salt bridge and give a more representative prediction of  $pK_a$ .

#### $\Delta$ +PHS L36D

The mutant L36D suffers from sampling problems, both of conformational and protonation space (Fig. 1). While at certain pH values CpHMD simulations correctly predict the  $pK_a$ , which experimentally is found to be 7.90,<sup>36</sup> there is no clear trend in  $pK_a$  prediction for simulations conducted at different levels of pH (Fig. 1). From visualization of the various trajectories, it is suggested that ASP36 may form a strong hydrogen bond with ASP21 in the MD simulations, stabilizing the deprotonated form. This scenario is seen at pH 4.5, where the simulation more accurately predicts a  $pK_a$  of 7.4. In other cases, ASP36 becomes buried in the hydrophobic interior of the protein, again leading to insufficient sampling of different protonation states. It is therefore likely that L36D needs to better sample conformational space in order to more effectively predict the  $pK_a$  of ASP36.

#### Analysis of CpHMD performance

It is clear the CpHMD method performs better at predicting the  $pK_a$  values of solvent-exposed residues, which possess  $pK_a$  values closer to their reference compounds (Table I). This is evident from the calculation of the root mean square error (RMSE) of predicted  $pK_a$  values, measured against the experimental work of García-Moreno to quantify this result, showing that residues on the surface of  $\Delta$ +PHS deviate from experiment with an RMSE of 1.23, whereas the RMSE for interior residues of the various  $\Delta$ +PHS mutants is 2.42 (Table II).<sup>31</sup> Most residues found at the surface of the protein encounter an increased number of transitions between protonated and deprotonated forms, and tend to converge relatively



**Figure 4**

RMSD as a function of MD time step for the  $\Delta$ +PHS mutant F34E protein at varying pH values.

**Table II**

RMS Errors of Predicted  $pK_a$  Values Against Experimental Values for Residues Located in Different Regions of the  $\Delta$ +PHS Protein (Exterior Residues) and  $\Delta$ +PHS Mutants (Interior Residues)<sup>31</sup>

	Surface	Interior
All residues	1.23	2.42
Aspartates	1.22	2.59
Glutamates	1.23	0.90
Lysines	–	3.12

Errors are computed with zero error if the predicted  $pK_a$  falls within the bounds of experimental  $pK_a$  with limiting values. Residues that do not incur transitions (ASP21 and L37D) are omitted from this calculation.

quickly ( $\sim 6$ – $8$  ns). The counterexample to this trend is ASP21, which likely fails to transition due to sampling problems derived from the persistence of its hydrogen bond with ASP19.

The importance of selecting a suitable pH range for titration and difficulties in achieving proper sampling of conformational space are illustrated in some  $pK_a$  predictions of interior residues for the various  $\Delta$ +PHS mutants. Given that many internal residues are found experimentally to have  $pK_a$  values shifted considerably from their reference  $pK_a$ , it is thus important to set up simulations over a wide pH range to conduct the titration. For example, in the case of L36D, simulations were performed at acidic pH under the assumption that the  $pK_a$  of the aspartic acid would exist closer to its reference value of 3.8. In fact, experimental results show this residue to titrate at a  $pK_a$  of 7.80. The selection of the pH range is also important for the stability of the system when performing CpHMD simulations, as illustrated by  $\Delta$ +PHS F34E.

Analyses of computed  $pK_a$  over time show internal residues to be far less converged compared to surface residues, making the prediction of accurate  $pK_a$  values more challenging. Although the CpHMD method applied in this study usually predicts the direction of the  $pK_a$  shift from the reference compounds correctly, there is still room for improvement in accurately predicting  $pK_a$  for internal residues.

Residues buried within the protein environment experience dielectric environments quite different from those at the surface of the protein, with their  $pK_a$  properties very susceptible to the nature of the residues in their vicinity and thus more difficult to treat computationally.<sup>31,42</sup> This difficulty is illustrated by the  $\Delta$ +PHS L36D and F34E proteins, where strong hydrogen bonds or salt bridges involving these titratable residues affect their protonation equilibria. Simulations of L36D do not contain any transitions between deprotonated and protonated forms owing to the persistence of an interaction between the ASP36 and ASP21 residues. For residues such as this, the use of an enhanced sampling method, such as accelerated MD, may assist in the sampling of relevant conformations and thus protonation states. The

requirement for increased sampling is also highlighted in instances where salt bridges persist throughout the simulation, as in the case of ARG81–GLU34 in the F34E mutant protein. The CpHMD method severely under-predicts the  $pK_a$  of this glutamate, suggesting it spends more time in its deprotonated form than experiment predicts.<sup>31</sup> It is possible, although not proven in our studies, that the strength of salt bridges sampled in our CpHMD method is overestimated under the GB implicit solvation, thus leading to error in predicting protonation state.<sup>51</sup>

Although not specifically quantified in this study, errors likely exist in CpHMD simulations due to the use of implicit solvation and conventional (non-polarizable) force fields. With regards to implicit solvation, issues regarding global protein movements would likely be dampened by the presence of explicit solvent molecules. Despite this generally accepted point, we believe CpHMD simulations employing implicit solvation still merit further study due to the simplicity of protonation changes and transition energy calculations. With regards to force fields, polarizable force fields would likely better capture the sensitivity of neighboring groups to changes in the protonation state. The topic of force field effects on constant pH MD simulations is further investigated by others in this issue.<sup>52</sup>

Although there exist problems with both implicit solvation and conventional force fields, the CpHMD method has been successful in predicting the  $pK_a$  of a significant number of residues from the test set from García-Moreno. This study has highlighted areas that may add significant improvement in the  $pK_a$  prediction capability of the method, such as enhanced conformational sampling and implementation of an improved solvation model. Future work will focus on testing other solvation models and the implementation of different accelerated molecular dynamics techniques, with the goal of achieving better sampling of physically meaningful conformations and protonation states.

## REFERENCES

1. Antosiewicz J, McCammon JA, Gilson MK. Prediction of pH-dependent properties of proteins. *J Mol Biol* 1994;238:415–436.
2. Bashford D, Karplus M.  $pK_a$ 's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* 1990;29:10219–10225.
3. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of  $pK_a$ s in proteins. *Proteins: Struct Funct Genet* 1993;15:252–265.
4. Mehler EL, Guarnieri F. A self-consistent microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys J* 1999;77:3–22.
5. Wisz MS, Hellinga HW. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. *Proteins: Struct Funct Genet* 2003;51:360–377.
6. Pokala N, Handel TM. Energy functions for protein design I: efficient and accurate continuum electrostatics and solvation. *Protein Sci* 2004;13:925–936.

7. Archontis G, Simonson T. Proton binding to proteins: a free-energy component analysis using a dielectric continuum model. *Biophys J* 2005;88:3888–3904.
8. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 1990;112:6127–6129.
9. Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculating  $pK_a$ s in proteins. *Biophys J* 2002;83:1731–1748.
10. Sham YY, Chu ZT, Warshel A. Consistent calculations of  $pK_a$ 's of ionizable residues in proteins: semi-microscopic and microscopic approaches. *J Phys Chem B* 1997;101:4458–4472.
11. Merz KM. Determination of  $pK_a$ s of ionizable groups in proteins: the  $pK_a$  of Glu 7 and 35 in hen egg white lysozyme and Glu 106 in human carbonic anhydrase II. *J Am Chem Soc* 1991;113:3572–3575.
12. Jensen JH, Li H, Robertson AD, Molina PA. Prediction and rationalization of protein  $pK_a$  values using QM and QM/MM methods. *J Phys Chem A* 2005;109:6634–6643.
13. Riccardi D, Schaefer P, Yang Y, Yu HB, Ghosh N, Prat-Resina X, König P, Li GH, Xu DG, Guo H, Elstner M, Cui Q. Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J Phys Chem B* 2006;110:6458–6469.
14. Mongan J, Case DA, McCammon JA. Constant pH molecular dynamics in generalized Born implicit solvent. *J Comput Chem* 2004;25:2038–2048.
15. Baptista AM, Teixeira VH, Soares CM. Constant-pH molecular dynamics using stochastic titration. *J Chem Phys* 2002;117:4184–4200.
16. Machuqueiro M, Baptista AM. Acidic range titration of HEWL using a constant-pH molecular dynamics method. *Proteins: Struct Funct Bioinf* 2008;72:289–298.
17. Dlugosz M, Antosiewicz JM. Constant-pH molecular dynamics simulations: a test case of succinic acid. *Chem Phys* 2004;302:161–170.
18. Dlugosz M, Antosiewicz JM, Robertson AD. Constant-pH molecular dynamics study of protonation-structure relationship in a heptapeptide derived from ovomucoid third domain. *Phys Rev E* 2004;69:021915.1–021915.10.
19. Baptista AM, Martel PJ, Petersen SB. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. *Proteins: Struct Funct Genet* 1997;27:523–544.
20. Borjesson U, Hunenberger PH. Explicit-solvent molecular dynamics simulation at constant pH: methodology and application to small amines. *J Chem Phys* 2001;114:9706–9719.
21. Borjesson U, Hunenberger PH. pH-dependent stability of a decalysine  $\alpha$ -helix studied by explicit-solvent molecular dynamics simulations at constant pH. *J Phys Chem B* 2004;108:13551–13559.
22. Lee MS, Salisbury FR, Brooks CL. Constant-pH molecular dynamics using continuous titration coordinates. *Proteins: Struct Funct Bioinf* 2004;56:738–752.
23. Khandogin J, Brooks CL. Constant pH molecular dynamics with proton tautomerism. *Biophys J* 2005;89:141–157.
24. Khandogin J, Chen JH, Brooks CL. Exploring atomistic details of pH-dependent peptide folding. *Proc Natl Acad Sci USA* 2006;103:18546–18550.
25. Williams SL, de Oliveira CAF, McCammon JA. Coupling constant pH molecular dynamics with accelerated molecular dynamics. *J Chem Theory Comput* 2010;6:560–568.
26. Meng YL, Roitberg AE. Constant pH replica exchange molecular dynamics in biomolecules using a discrete protonation model. *J Chem Theory Comput* 2010;6:1401–1412.
27. Mongan J, Case DA. Biomolecular simulations at constant pH. *Curr Opin Struct Biol* 2005;15:157–163.
28. Chen JH, Brooks CL, Khandogin J. Recent advances in implicit solvent-based methods for biomolecular simulations. *Curr Opin Struct Biol* 2008;18:140–148.
29. Baker NA, Bashford D, Case DA. Implicit solvent electrostatics in biomolecular simulation. In: Leimkuhler B, Chipot C, Elber R, Laaksonen A, Mark A, Schlick T, Schütte C, Skeel R, editors. *New algorithms for macromolecular simulation*, Vol. 49. New York: Springer; 2006. pp. 263–295.
30. <http://amylase.ucd.ie/pKacoop/>.
31. Castaneda CA, Fitch CA, Majumdar A, Khangulov V, Schlessman JL, Garcia-Moreno BE. Molecular determinants of the  $pK_a$  values of Asp and Glu residues in staphylococcal nuclease. *Proteins: Struct Funct Bioinf* 2009;77:570–588.
32. Isom DG, Cannon BR, Castaneda CA, Robinson A, Garcia-Moreno BE. High tolerance for ionizable residues in the hydrophobic interior of proteins. *Proc Natl Acad Sci USA* 2008;105:17784–17788.
33. Isom DG, Castaneda CA, Cannon BR, Velu PD, Garcia-Moreno BE. Charges in the hydrophobic interior of proteins. *Proc Natl Acad Sci USA* 2010;107:16096–16100.
34. Isom DG, Castaneda CA, Cannon BR, Garcia-Moreno BE. Large shifts in  $pK(a)$  values of lysine residues buried inside a protein. *Proc Natl Acad Sci USA* 2011;108:5260–5265.
35. Harms MJ, Schlessman JL, Sue GR, Garcia-Moreno BE. In press.
36. Cannon BR, Isom DG, Garcia-Moreno BE. In press.
37. Chimenti MS, Khangulov VS, Robinson AC, Heroux A, Majumdar A, Schlessman JL, Garcia-Moreno BE. In press.
38. Bashford D, Case DA, Dalvit C, Tennant L, Wright PE. Electrostatic calculations of side-chain  $pK_a$  values in myoglobin and comparison with NMR data for histidines. *Biochemistry* 1993;32:8045–8056.
39. Kyte J. *Structure in protein chemistry*. New York: Garland Publishing Inc.; 1995. p.64.
40. Onufriev A, Case DA, Ullmann GM. A novel view of pH titration in biomolecules. *Biochemistry* 2001;40:3413–3419.
41. Cantor CR, Schimmel PR. *Biophysical chemistry. Part III. The behavior of biological macromolecules*. San Francisco: W.H. Freeman and Co.; 1980. pp. 863–866.
42. Fitch CA, Whitten ST, Hilser VJ, Garcia-Moreno BE. Molecular mechanisms of pH-driven conformational transitions of proteins: insights from continuum electrostatics calculations of acid unfolding. *Proteins: Struct Funct Bioinf* 2006;63:113–126.
43. Wang JM, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* 2000;21:1049–1074.
44. Onufriev A, Case DA, Bashford D. Effective Born radii in the generalized Born approximation: the importance of being perfect. *J Comput Chem* 2002;23:1297–1304.
45. Onufriev A, Bashford D, Case DA. Modification of the generalized Born model suitable for macromolecules. *J Phys Chem B* 2000;104:3712–3720.
46. Onufriev A, Bashford D, Case DA. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins: Struct Funct Bioinf* 2004;55:383–394.
47. Ryckaert JP, Ciccotti G, Berendsen HJC. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys* 1977;23:327–341.
48. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys* 1984;81:3684–3690.
49. MATLAB, Version 7.11.0. Natick, Massachusetts: The MathWorks Inc.; 2010.
50. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein  $pK_a$  values. *Proteins: Struct Funct Bioinf* 2005;61:704–721.
51. Chen J, Im W, Brooks CL. Balancing solvation and intramolecular interactions: toward a consistent generalized Born force field. *J Am Chem Soc* 2006;128:3728–3736.
52. Machuqueiro M, Baptista AM. Is the prediction of  $pK_a$  values by constant-pH molecular dynamics being hindered by inherited problems? *Proteins: Struct Funct Bioinf* 2011;79:3437–3447.