Predicting the Acid/Base Behavior of Proteins: A Constant-pH Monte Carlo Approach with Generalized Born Solvent

Alexey Aleksandrov,[†] Savvas Polydorides,[‡] Georgios Archontis,*[,‡] and Thomas Simonson*[,†]

Laboratoire de Biochimie (CNRS UMR7654), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France, and Department of Physics, University of Cyprus, PO20537, CY1678, Nicosia, Cyprus

Received: May 14, 2010; Revised Manuscript Received: July 7, 2010

The acid/base properties of proteins are essential in biochemistry, and proton binding is a valuable reporter on electrostatic interactions. We propose a constant-pH Monte Carlo strategy to compute protonation free energies and pK_a's. The solvent is described implicitly, through a generalized Born model. The electronic polarizability and backbone motions of the protein are included through the protein dielectric constant. Side chain motions are described explicitly, by the Monte Carlo scheme. An efficient computational algorithm is described, which allows us to treat the fluctuating shape of the protein/solvent boundary in a way that is numerically exact (within the GB framework); this contrasts with several previous constant-pH approaches. For a test set of six proteins and 78 titratable groups, the model performs well, with an rms error of 1.2 pH units. While this is slightly greater than a simple Null model (rms error of 1.1) and a fully empirical model (rms error of 0.9), it is obtained using physically meaningful model parameters, including a low protein dielectric of four. Importantly, similar performance is obtained for side chains with large and small pK_a shifts (relative to a standard model compound). The titration curve slopes and the conformations sampled are reasonable. Several directions to improve the method further are discussed.

1. Introduction

The acid/base properties of proteins are essential in biochemistry and have been studied for almost a century.[1−3] Proton transfer is central to both respiration and photosynthesis;[4] many enzyme reactions include proton binding steps,[5] while protein folding, supramolecular assembly, adsorption, and binding to ligands are all sensitive to pH.[6,7] In addition, protons can be valuable reporters on electrostatic interactions and dielectric relaxation,[8−10] which have even broader significance.[9−13] Experimental methods, especially high-resolution structure determination, have opened the way to a thorough understanding of acid/base biochemistry and electrostatic interactions in general. However, the complexity of proteins and their aqueous environment are such that biophysical models are almost always needed to complement and interpret the experiments: X-ray structures do not reveal electric fields.

Simple, empirical models have been developed in recent years that often give good accuracy (around one pK_a unit) and can help quantify the main effects.[14,15] For example, the popular PROPKA program[14] takes into account hydrogen bonds involving the titratable side chains, the extent of solvation/desolvation of these side chains, and their proximity to ionized groups. However, detailed physical theories obviously provide a deeper understanding. For example, Hush and Marcus used a two-step thought experiment to introduce a new charge, such as a proton or electron, allowing them to isolate the free energy contributions of electronic and dipolar reorganization.[16−19] Similarly, dielectric continuum theory quantifies the desolvation of a charge through its "self-energy",[20−23] which has a simple relation to the reorganization free energy for introducing the charge.[24] Thus, considerable efforts have been made to develop physically meaningful simulation models that give reasonable accuracy for pK_a's and can at the same time increase our understanding of protein electrostatics.

The most detailed and accurate models use an atomistic treatment of both the protein and the surrounding solvent, combined with molecular dynamics (MD) or Monte Carlo (MC) sampling of conformational space.[25−28] Fully quantum mechanical treatments are still not practical for most problems, but mixed quantum/classical treatments are common for studying enzyme reactions.[27,28,13,29] While fully atomistic models have been applied to proton binding and transfer,[30−36] there are many problems where they cannot be routinely used. In particular, when a protein is studied over a broad pH range, where many groups can bind or release protons, additional approximations are necessary. The oldest and simplest approach treats the protein and its surroundings as two homogeneous, isotropic, dielectric media and solves the Poisson−Boltzmann equation to obtain free energies, sometimes analytically,[20,37,38] but usually numerically.[39,23,40,41] The simple, two-dielectric approach has very severe limitations, discussed in the literature and in the Discussion. More recent models use a hybrid approach, where part of the system and/or some of its degrees of freedom are treated explicitly and atomistically, while other parts or degrees of freedom are "integrated out" and treated implicitly.[42] The implicit treatment often relies on a dielectric continuum model.[20,39,23,43,41,42,44,45,10]

Two of the most important, recent, hybrid approaches are the "multiconformation" Poisson−Boltzmann, or MCPB, methods[46−49] and the so-called "constant-pH" MD methods.[50−53] With MCPB, the protein and solvent are treated as two dielectric media, while protein side chain conformations are explored with Monte Carlo. The protein backbone is held fixed, and its conformational flexibility is absorbed into the protein dielectric constant.[54] With constant-pH MD, the protein and its motions

* Corresponding author. E-mail: thomas.simonson@polytechnique.fr; archonti@ucy.ac.cy.
[†] Ecole Polytechnique.
[‡] University of Cyprus.

10.1021/jp104406x
Published on Web 07/26/2010

are described in full atomistic detail, while the solvent is modeled as a dielectric continuum. The motions are sampled by MD simulation, and proton binding/release is treated through an extended ensemble that mimics a semigrand canonical ensemble (see Theory section). The atomic charges in the protein usually have fixed magnitudes, which implies that the electronic polarizability of the protein is absorbed into the protein dielectric constant. Since MD simulations with a PB solvent are rather expensive,[43,55,56] constant-pH MD is often done with a generalized Born (GB) solvent model,[57−62] which contains similar physics but is more efficient.[52,53]

Here, we present a novel hybrid approach. As with MCPB, we use a Monte Carlo method to explore the conformations of all protein side chains, and we hold the protein backbone fixed; its motions are treated implicitly, through the protein dielectric constant. However, our method has two distinctive features. First, we use a generalized Born solvent, as opposed to the PB solvent used with MCPB.[46−49] Second, we accurately account for changes in the shape of the protein volume (and the remaining solvent volume), which occur constantly as a result of the protein side chain fluctuations. Until now, only the constant-pH approach allowed an exact description of these volume changes. All other hybrid approaches require an additional approximation; usually, an average protein−solvent boundary is used, possibly with additional, ad hoc corrections.[46−49,63−65] Even though these corrections appear quite successful in several published applications,[63−65] it is of great interest to derive an essentially exact scheme and eliminate one source of empiricism from the model. This is done here in an efficient way by use of a recent, "residue-GB" variant, which has a "pairwise decomposable" property:[66] side chain−side chain and side chain−backbone interactions can be precomputed and tabulated, following a method introduced for computational protein design.[66,67] We show in the Theory section how this is possible without loss of accuracy, even though the generalized Born interaction energy is a many-body function.[57−59]

The method is tested on six small proteins, including a total of 78 titratable groups. The accuracy is reasonable: only slightly poorer than the best empirical and "physics-based" methods and comparable to several other physics-based methods. The best results are obtained with a model parametrization that is physically reasonable. The performance is comparable for side chains with unshifted and with strongly shifted pK_a values (shifted compared to standard model compounds). There are several clear directions to improve the method systematically. It is straightforward to extend it to compute redox potentials and to treat processes where proton and electron binding are coupled. It should also be possible to apply the method with a PB solvent instead of a GB solvent.

The article is organized as follows. In the next, Theory, section, we present the semigrand canonical ensemble, the classical mechanical treatment of proton binding, the residue-GB model, and the algorithm that reduces the computational complexity to a quadratic dependence on protein size. Computational details are described next. In the Results, we begin by presenting a systematic comparison between the residue-GB model and a standard PB model. This is appropriate since this work is the first application of residue-GB. Next, we present pK_a calculations for our test set of proteins. The last section is a Discussion. In particular, we discuss the limitations of continuum models, and the meaning of the protein dielectric constant in the context of hybrid models, where some of the relaxation channels are modeled explicitly and others are modeled implicitly.

2. Theory: Constant-pH Monte Carlo with Generalized Born Solvent

2.1. Statistical Mechanical Framework. Several authors have described the framework for constant-pH simulations using MD or MC.[50−53,48] A similar framework has been described and used for simulations with variable numbers of water molecules.[68−70] Following Baptista et al.,[50] we consider a dilute solution of the protein of interest, with a constant volume V and temperature T, closed with respect to protein and water but open with respect to protons. The corresponding thermodynamic ensemble is a slight variation of the grand canonical ensemble;[71] the partition function has the form (e.g., see eq 1-60 in ref 71)

$$\Xi(V,T,\mu) = \sum_{j,N} \exp(-\beta E_j(N,V))\exp(\beta N\mu) \qquad (1)$$

where μ is the chemical potential of the proton; the sum is over the number of protons N and the different states j of the system, which are characterized by their energies E_j; $\beta = 1/kT$; and k is Boltzmann's constant. For convenience, we have assumed the energy states are discrete; this would be the case for a finite quantum mechanical system, and it will also be the case in the applications below (since we will use a fixed-backbone, side chain rotamer description of the protein's conformational space, along with an implicit solvent model). The statistical probability $P_j(N; V, T, \mu)$ of a particular state can be written

$$P_j(N;V,T,\mu) = \exp(-\beta E_j(N,V))\exp(\beta N\mu)/\Xi(V,T,\mu) \qquad (2)$$

We will not consider the solvent explicitly; rather, we integrate over the solvent degrees of freedom.[72,42,73] The probability function keeps the same form, but we must replace the energy E_j by a potential of mean force, or PMF, denoted W_J[42]

$$P_J(N;V,T,\mu) = \frac{\exp(-\beta W_J(N;V))\exp(\beta N\mu)}{\sum_{J,N} \exp(-\beta W_J(N;V))\exp(\beta N\mu)} \qquad (3)$$

The index J now refers to an energy state of the protein alone (bathed in implicit solvent). Below, to approximate the PMF, we will use a continuum dielectric treatment, specifically, a generalized Born variant.[57,42,61,44] For the Monte Carlo simulations, below, we do not need to sample the protein velocities; therefore, we also integrate them out. This does not change the form of the probability $P_J(N; V, T, \mu)$; however, J now refers to a *conformational* state of the protein. Notice that J also identifies the specific location of all the protons bound to the protein. If we compare two states that differ by the addition of a proton to a specific titratable side chain, with the protein in a given conformational state J, the ratio of Boltzmann probabilities has the form

$$\frac{P_J(N+1)}{P_J(N)} = e^{-\beta(W_J(N+1)-W_J(N))}e^{\beta\mu} = e^{-\beta(W_J(N+1)-W_J(N))-2.303\mathrm{pH}+\beta\mu^0} \qquad (4)$$

The last equality uses the relation between μ and the pH: $\beta\mu = \beta\mu^0 - 2.303\mathrm{pH}$, where μ^0 is the proton chemical potential in aqueous solution in the standard state. For simplicity, in eq 4,

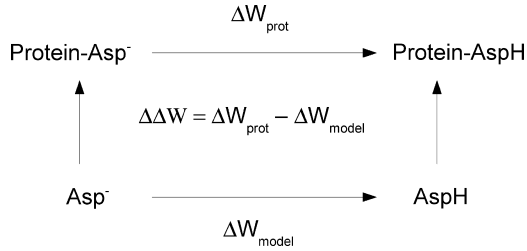

Figure 1. Thermodynamic cycle to analyze pK_a shifts: example of an Asp side chain. The upper leg represents proton binding to an Asp side chain in a protein. The lower leg represents proton binding to a model compound in solution. The double PMF difference between the two legs is $\Delta\Delta W = \Delta W_{prot} - \Delta W_{model}$. All four states are assumed to be at standard state concentrations (ideal solutions with 1 M concentrations).

we have kept the same index J for the two states, even though the $N + 1$ state has an extra proton; this slightly abusive notation helps emphasize that the protein conformation is otherwise unchanged, and only a proton has been added to a particular location. The Boltzmann probability distribution in eqs 3 and 4 will be sampled using the standard, Metropolis, Monte Carlo algorithm[74,75] (see below).

2.2. Classical Mechanical Treatment of Proton Binding. Proton binding is described within a classical mechanical, molecular mechanics framework, first proposed by Warshel and co-workers,[30] and abundantly used ever since.[23,40,41,76,33] The titrating protons, like the other atoms in the system, are treated as classical mechanical particles, bearing a partial charge and interacting with the other atoms through Coulombic terms, stereochemical terms, and van der Waals terms. The force field used here[77] employs fixed partial charges; i.e., electronic polarizability is treated in a simple, mean-field way.

This approach would not be appropriate to compute absolute protonation free energies; rather, one uses it to compute the difference between the PMF change for protonation of a protein side chain and that of a suitable model compound in solution.[30] The model compound is typically an analogue of the corresponding side chain; for example, aspartate with neutral backbone blocking groups (2*N*-acetyl-1*N*-methylaspartic acid-1-amide) could serve as a model compound for Asp side chain protonation (see below). We consider the thermodynamic cycle in Figure 1: subtracting the PMF change for the upper and lower legs yields a double PMF difference, $\Delta\Delta W$. We expect that taking this difference leads to a cancellation of most of the error associated with the simple, molecular mechanical treatment employed.[30,23,40,41,76,33] To obtain an absolute PMF change for protonation/deprotonation in the protein, $W_J(N + 1) - W_J(N)$ (eq 4), we must then add back the contribution of the model compound in solution. This is straightforward since the model compound's standard protonation free energy ΔG_{model} at a given pH has a simple relation to its pK_a and to the PMF change for the model compound

$$\beta\Delta G_{model} = \beta\Delta W_{model} - \beta\mu^0 = -2.303\mathrm{p}K_{a,model} \quad (5)$$

where p$K_{a,model}$ is the pK_a of the model compound in aqueous solution. The μ^0 term appears because with our definition of the PMF ΔW_{model} does not include any contributions from the solvated proton. In eq 5, we assume the states linked by the thermodynamic cycle of Figure 1 correspond to standard state concentrations.

2.3. Reducing the Computational Complexity: Residue-GB. To make the Monte Carlo calculations efficient, we use a strategy borrowed from the field of computational protein design.[67,78,79] The interaction energies are computed ahead of time for all side chain pairs in the protein, allowing for all possible rotamers and protonation states. The Monte Carlo exploration of rotamers and protonation states can then be done very efficiently, with the interaction energies obtained from lookup tables.

At first glance, this strategy appears impossible with a continuum dielectric solvent model. Indeed, in continuum electrostatics, the effective interaction between two residues depends on the entire protein's shape and the complementary volume occupied by high dielectric solvent.[22] Therefore, continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[10,63] Here, we overcome this difficulty thanks to a novel generalized Born model that is residue-pairwise and can be used efficiently for constant pH simulations, as well as for protein design.[66] Two steps make the scheme pairwise. (i) First, we adopt an expression for the interaction energy between two residues R and R′ that depends on the product $B = B_R B_{R'}$ of their *residue* Born solvation radii. These radii reflect the desolvation, or burial, within the protein of each residue. With most GB models, they are readily obtained from residue-pairwise quantities. (ii) Second, we fit the RR′ interaction energy by a simple function of B; the fitting coefficients depend only on the pair RR′, *not* on its environment. In effect, the quantity B captures all the information that is relevant about the pair's dielectric environment. The numerical accuracy of the fitting scheme can be made arbitrarily high; the fitting scheme chosen below can be considered essentially exact.[66] Below, we describe steps (i) and (ii) in detail.

2.3.1. Step (i): Residue Generalized Born. With GB, the electrostatic energy includes both a direct, Coulomb term and a contribution from the solvent, polarized by the solute charges. Treating the solvent as a linear, homogeneous, dielectric medium, the total electrostatic energy has the form

$$\begin{aligned} E^{elec} &= E^{Coul} + \Delta G^{solv} \\ &= \frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{\epsilon_p r_{ij}} + \frac{1}{2}\sum_{ij} g_{ij} \end{aligned} \quad (6)$$

where the sums are over all pairs of protein charges and the second sum includes diagonal terms, $i = j$. In eq 6, ε_p is the protein dielectric constant, which is usually set to one in constant-pH studies but will be set to 4 or 8 in the application below (to account for the backbone and electronic degrees of freedom). The second sum, ΔG^{solv}, represents the electrostatic solvation free energy of the protein (in the given conformation).[42] The term g_{ij} represents the interaction between a protein charge q_i and the solvent polarization induced by another charge, q_j. We refer to it as a GB interaction or screening energy. In the standard, Atomic GB model,[57] this term is approximated by

$$g_{ij} = g(\underline{r}_i, \underline{r}_j) = \frac{\tau q_i q_j}{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}} \quad (7)$$

where $r_{ij} = |\underline{r}_i - \underline{r}_j|$; $\tau = 1/\varepsilon_w - 1/\varepsilon_p$; ε_w is the solvent dielectric constant (80 at room temperature); and b_i and b_j are

effective, *atomic*, "solvation radii" of the charges i, j. The interaction between two residues, R and R′, can then be written

$$g_{RR'} = \sum_{i\in R, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}} \quad (8)$$

Earlier, we obtained a Residue GB model by replacing the solvation radii b_i within a particular residue R by an average value b_R.[66] The interaction between two residues R and R′ can then be written

$$g_{RR'} = \sum_{i\in R, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + B_R B_{R'} \exp[-r_{ij}^2/4B_R B_{R'}])^{1/2}} \quad (9)$$

The average solvation radius B_R is related to the GB self-energy of residue R[58,59]

$$E_R^{self} = \sum_{i\in R} E_i^{self} = \frac{\tau}{2}\sum_{i\in R}\frac{q_i^2}{b_i} \stackrel{def}{=} \frac{\tau}{2}\sum_{i\in R}\frac{q_i^2}{B_R} \quad (10)$$

Equivalently

$$(\sum_{i\in R} q_i^2)\frac{1}{B_R} = \sum_{i\in R}\frac{q_i^2}{b_i} \quad (11)$$

Thus, B_R is a harmonic average over the b_i, $i \in R$, weighted by the squared charges.[66]

2.3.2. Step (ii): The Residue Self-Energies Are Sufficient to Parametrize the Pair Screening Energies. A residue-pairwise scheme can now be devised. We note that, for fixed interatomic distances r_{ij}, $g_{RR'}$ is a slowly varying function of $B = B_R B_{R'}$. This dependency can be approximated by a low-order polynomial

$$g(B;r) = (r^2 + B\exp[-r^2/4B])^{-1/2} \approx c_1(r) + c_2(r)B + c_3(r)B^2 + O(B^3) \quad (12)$$

The interaction energy $g_{RR'}$ then takes the form

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 \quad (13)$$

Although this approximation usually holds for a large range of $B = B_R B_{R'}$ values, we actually prefer to use a more complex fitting function,[66] which provides very high accuracy for all residue pairs, all rotamer combinations, and all relevant B values

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 + c_4^{RR'}B^{-1/2} + c_5^{RR'}B^{-3/2} \quad (14)$$

The two rightmost terms improve the fit accuracy for large B values. The coefficients $c_i^{RR'}$, $i = 1, ..., 5$, will be precomputed for all residue pairs, allowing for all combinations of rotamers and protonation states. During the Monte Carlo simulation, we can then obtain $B_R B_{R'}$ and hence $g_{RR'}$ very efficiently, on-the-fly.

3. Computational Details

3.1. Computation of Energy Matrices. We first consider individual side chains and compute their interactions with the backbone, including all possible titration states and rotamers (see below). Backbone atoms are held fixed in their experimental positions. Pro, Ala, and Gly are treated as a part of the backbone, as are cysteines involved in disulfide bridges. The side chain is positioned in a particular rotamer, then the atomic positions are slightly optimized, through 25 steps of Powell energy minimization, only considering interactions with the protein backbone. The side chain−backbone interaction energy is then computed and stored, and the side chain's "minimized-rotamer" coordinates are saved for future use.

For a side chain pair, (R, R′), we again consider all choices of titration states and rotamers. We use the minimized-rotamer coordinates obtained above; the pair of side chains is further minimized through 10 steps of Powell minimization, to further alleviate the steric overlap that can result from the rotamer approximation. Interactions between the pair and with the backbone are included in the minimization, and the backbone is held fixed as before. The side chain interaction energy is computed and stored. Individual energy terms are stored separately; for example, the contribution of side chain R to the GB self-energy of residue R′ and the contribution of R′ to the GB self-energy of R are stored in a square, asymmetric, self-energy matrix. All the rotamer construction and energy calculations are done with the Xplor program.[80]

3.2. Five-Point Fitting Procedure for the GB Interaction Energies. For a particular pair of side chains, R and R′, and choice of rotamers, the GB interaction energy $g_{RR'}$ was computed for 20 B values, evenly spaced between 1 and 150 Å^2. These energies were fit by the five-point function given in eq 14. The fitting was done with a Fortran program based on the general linear fit subroutine LFIT from Numerical Recipes.[81,66]

3.3. Monte Carlo Simulation Protocol. The Monte Carlo simulation explores the space of side chain rotamers and titration states. Asp, Glu, His, Lys, and Tyr are all considered titratable; there are no cysteines in our data set, except those involved in disulfide bonds. Histidines have three possible protonation states: doubly protonated and singly protonated on either Nε or Nδ. The backbone N- and C-termini are also titratable.

Each Monte Carlo move $(J, N) \rightarrow (J', N')$ changes either the protonation state or the rotamer of either a single group or a pair of groups. The first group is chosen randomly; half of the time, a second group is chosen randomly from among those that have a strong interaction with the first one (±2 kcal/mol or more). Moves are accepted or rejected according to the Metropolis algorithm. If the new state (J', N') has an increased Boltzmann probability, the move is accepted. If it has a decreased probability, the move is accepted with the probability

$$p_{(J,N)\rightarrow(J',N')} = \exp(-\beta\Delta W + \beta(N' - N)\mu) \quad (15)$$

where $\beta = 1/kT$; k is Boltzmann's constant; T is the temperature; and $\Delta W = W_{J'}(N') - W_J(N)$ is the PMF change. For a move that adds a proton (eq 4), the PMF change is computed with the help of the thermodynamic cycle in Figure 1, and the acceptance probability has the form

$$p_{(J,N)\rightarrow(J,N+1)} = e^{-\beta(W_J(N+1)-W_J(N))-2.303pH+\beta\mu^0} = e^{-\beta\Delta\Delta W+2.303(pK_{a,model}-pH)} \quad (16)$$

TABLE 1: Model pK_a Values Used in This Work

titrating group	model pK_a
Asp	4.0
Glu	4.4
Tyr	10.3
Lys	10.4
C-Ter	3.2
N-Ter	9.1
His_δ	6.6
His_ε	7.0

The last equality has made use of eqs 3 and 5. For a move that removes a proton or adds or removes two protons, the acceptance probability is obtained similarly.

Simulations are done at a fixed pH, at a temperature of $T = 300$ K. Each simulation is started from a set of randomly chosen rotamers and titration states. Equilibration is done for 1 million steps. At this point, rotamers that are never accessed are removed, and a further 5 million steps are done. The pH is varied from 0 to 14 in steps of 0.5 units. Thus, a complete pH scan involves 174 million MC steps.

3.4. Model Compounds. The protonation free energy of a residue in the protein is computed as the sum of the protonation free energy of a model compound in solvent and a contribution from the protein, $\Delta\Delta W$, defined in Figure 1.[23,30] The protonation free energy of the model compound in solvent at a given pH is given by eq 5. The model compounds and their pK_a's are listed in Table 1. For Asp, Glu, His, Lys, and Tyr, the model compound is represented by the side chain and backbone atoms of the particular residue. For the N-terminus (respectively, C-terminus), we use the backbone of the first (last) two residues.

3.5. Rotamer Library. Side chain dihedral angles were taken from the Tuffery rotamer library,[82] with the following changes. For C-ter, Asp, and Glu, the number of rotamers was doubled, to ensure that alternate orientations of protonated carboxylic acids were present. The number of Tyr rotamers was also doubled, so that for each Tuffery rotamer we allow both possible orientations of the hydroxyl hydrogen in-plane with the phenyl ring. Similarly, for Ser and Thr, the number of rotamers was tripled to allow, for each Tuffery rotamer, three orientations of the hydroxyl hydrogen. For each side chain, we also added between one and three "native" rotamers, which combine the experimental position of the side chain heavy atoms with the hydrogen orientations just described (two per Asp, Glu, C-ter, and Tyr, three per Ser and Thr).

3.6. Force Field and GB Parameters. We use the Amber all-atom energy function, version parm11.[77] For the titratable residues, the charges of backbone atoms (N, HN, C_α, H_α, C, and O) were set to have average Amber values for these atoms. Charges for side chain atoms of titratable residues were adopted with slight modifications from the Amber force field. No cutoff was used for the nonbonded interactions.

For the solvent, we use a generalized Born variant developed to be consistent with the Amber force field.[60,83] It uses the self-energy treatment of Hawkins et al.,[58] so we refer to it as the GB/HCT model. Recently, we optimized the atomic volumes and other GB parameters to maximize agreement with a large set of Poisson−Boltzmann reference calculations, including mutation free energies and conformational free energies, all done with the same Amber force field, version parm11.[84] The PB reference calculations used atomic radii specifically optimized for use with the Amber charges.[85] The force field and GB variant are implemented in the Xplor program,[80,86,83] which was used for all the energy calculations.

TABLE 2: Protein Structures Used in This Work

protein	PDB structure	X-ray resolution	experimental pH
bovine pancreatic trypsin inhibitor (BPTI)	4PTI[104]	1.50 Å	NA[a]
streptococcal protein G, B1 Ig-binding domain	1PGA[144]	2.07 Å	4.5
turkey ovomucoid third domain (OMTKY3)	2GKR[145]	1.16 Å	7.5
chicken lysozyme	2LZT[146]	1.97 Å	4.5
Bacillus amyloliquefaciens barnase	1A2P[147]	1.50 Å	7.5
Escherichia coli thioredoxin, oxidized	2TRX[148]	1.68 Å	7.5

[a] Not available.

3.7. Protein Set and Atomic Coordinates. The pK_a calculations were done for six proteins, listed in Table 2. Their chain lengths are between 56 and 129 amino acids. Crystal structures were used, with a resolution of 2.1 Å or better. Crystal waters and ligands were removed. Hydrogens of protonated carboxylic acids (aspartates, glutamates, C-termini) were added in the most common, syn orientation. Other hydrogens were first positioned with the HBUILD facility in XPlor;[87] all hydrogen positions were then adjusted by full-energy minimization, with all non-hydrogen atoms fixed and all titratable groups kept in their protonated forms.

3.8. Poisson−Boltzmann pK_a Calculations. For comparison to the GB method, single-conformation Poisson−Boltzmann (SC-PB) calculations were done, using the MEAD program[88,89] to solve the linearized PB equation. We used the slightly modified Amber atomic charges described above, along with atomic radii specifically optimized for PB calculations with the Amber charges.[85] The protein dielectric constant was set to 4 or 8, and the solvent dielectric constant was set to 80. The boundary between the protein and solvent regions was taken to be the molecular surface, constructed with a probe radius of 1.4 Å. Calculations were done at 300 K with an ionic strength of 0.150 M. The PB equation was solved using a three-step focusing procedure, and the spacings between grid points at the three steps were 4, 1, and 0.5 Å. To determine the protonation pattern at a given pH, we used the Monte Carlo procedure implemented in the MCTI program,[90] obtaining the occupancies of the deprotonated and protonated states of all the titratable residues. The pH was varied from −10 to 30 in steps of 0.1 pH unit. The pK_a of each titratable group was taken to be the pH where the occupancies of the protonated and deprotonated forms were equal.

3.9. PROPKA Calculations. We used the PROPKA interactive web interface, http://propka.ki.ku.dk/, which takes a PDB code as input and returns the pK_a values for titratable amino acids.

3.10. Comparing Residue-GB to Atomic-GB and to the Poisson Model. To test the residue-GB method, we report a series of tests, comparing it to atomic-GB and the Poisson model. We compute free energy changes associated with conformational, protonation, and chemical modifications in four medium-sized proteins. The chemical modifications are described in the Supporting Information.

3.10.1. Rotamer Calculations. For each protein, we constructed a set of 3000 structures by randomizing the side chain rotamers of all residues except prolines, alanines, and cysteines. Small voids in the interior of the rotameric structures were filled by dummy atoms, to prevent the occurrence of artificial, high-

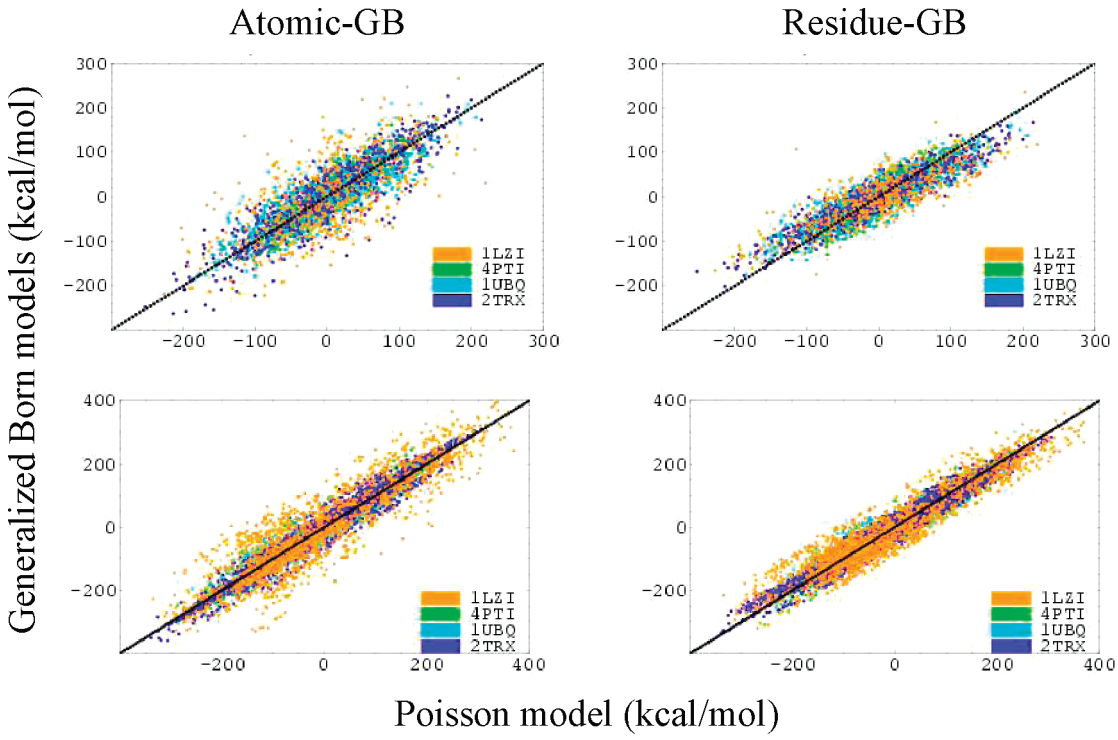

Figure 2. Solvation free energies of four proteins for multiple rotamer combinations (upper panels) and multiple protonation states (lower panels). Atomic-GB (left, vertical axis) and residue-GB (right) are compared to the Poisson equation (horizontal axis). The proteins are identified by their PDB codes and a color scheme.

dielectric internal cavities. The protein backbones were held fixed in their X-ray conformations, taken from the following four PDB structures: 1UBQ for ubiquitin; 4PTI for BPTI; 1LZ1 for lysozyme; 2TRX for thioredoxin.

3.10.2. Mutations in the Protonation State of Titratable Residues. In each of the four proteins, the titratable residues Asp, Glu, His, Cys, and Lys were initially assigned their most common protonation states at physiological pH. Subsequently, the charge state of one titratable residue at a time was modified, and the solvation free energy change was computed for 100 protein conformations, created by randomizing the rotameric state of all side chains. The chemical types, charges, and radii for the different charge states of titratable residues were taken from the AMBER force field, as for the Monte Carlo simulations (see above).

3.10.3. Reference Calculations with the Poisson Model. For these calculations, where residue-GB is compared to the Poisson model, we used a slightly different Poisson protocol, compared to the pK_a calculations, above. For a given protein structure, we solved the Poisson equation with the finite-difference program UHBD[91] (instead of MEAD, above). The protein/solvent dielectric boundary was defined by the molecular surface of the protein. The solution employed a two-step focusing procedure and a cubic grid with spacings of 0.8 and 0.4 Å. All calculations were done with a protein dielectric constant $\varepsilon_p = 1$ and a solvent dielectric $\varepsilon_s = 80$. The same protein dielectric is used for the GB model (unlike the pK_a case, above). Indeed, for these GB/PB comparisons, what is essential is the consistency between the various treatments, rather than the precise choice of protein dielectric constant. The molecular surface was constructed with 2000 points per atom, using a probe sphere of radius 2 Å and the boundary smoothing method in UHBD.[91] The atomic charges corresponded to the AMBER, all-atom force field (see above). The atomic radii were those specifically optimized for Poisson calculations with the AMBER charges.[85]

4. Results

4.1. Residue-GB Compares Well with the Poisson Model. This work represents the first application of residue-GB. Therefore, before describing the pK_a results, we report a series of tests that specifically compare residue-GB to the more traditional, atomic-GB and to the Poisson method. We introduced a large number of conformational, protonation, and chemical modifications into four, medium-sized proteins, ubiquitin, BPTI, lysozyme, and thioredoxin, and computed the corresponding free energy changes. The Poisson model is considered the benchmark for accuracy, following common practice with GB model development. Below, we describe the conformational and titration changes, and the chemical mutations are described in Supporting Information.

4.1.1. Conformational Energies. The GB electrostatic solvation free energies for the rotameric structures of the four proteins lysozyme, thioredoxin, ubiquitin, and BPTI are plotted against the corresponding PE values in Figure 2. Results are shown for both atomic- and residue-GB. The free energies vary over a 500 kcal/mol range. The present variant of atomic GB was optimized earlier by comparison to PE[84] and agrees very well with the PE model across the whole energy range. The rms differences between the GB and PE electrostatic solvation free energies are given in Table 3. The values for the four proteins vary between 21 and 45 kcal/mol for atomic-GB and between 21 and 40 kcal/mol for residue-GB. Overall, residue-GB performs as well as atomic-GB for three of the proteins and slightly better for lysozyme.

4.1.2. Protonation of Titratable Residues. Here, we systematically change the protonation state of titratable amino acids in the four proteins considered above. The resulting GB solvation energies are plotted against the corresponding PE energies in Figure 2. The corresponding rms differences are given in Table 3. The GB values correlate well with PE, with

TABLE 3: Comparing Atomic-GB and Residue-GB to the Poisson Model[a]

	rotamers		protonation states	
protein	atomic-GB	residue-GB	atomic-GB	residue-GB
1LZ1	45.3	39.4	52.0	42.8
2TRX	30.2	28.9	24.5	24.9
1UBQ	26.0	27.7	25.4	24.9
4PTI	20.7	20.9	22.2	22.7

[a] RMS deviation (kcal/mol) between the solvation free energies from GB and the Poisson model (PE) for random rotamer combinations and protonation states of titratable side chains in four proteins (indicated by their PDB code; see text).

rms differences of between 22 and 52 kcal/mol with atomic-GB for the four proteins and between 23 and 43 kcal/mol for residue-GB. Residue-GB is somewhat better than atomic-GB for lysozyme and comparable for the other three proteins. All the GB pK_a results, below, were obtained with residue-GB.

4.2. Comparison between Computed and Experimental pK_a's. ***4.2.1. Overview of the Results.*** We now consider the titration behavior of 78 titratable groups in six proteins. Our MC simulations yield the distribution of protonation states for each group as a function of pH, in other words, the titration curves. Experiments usually yield the same information[3] (typically deduced from a change in the NMR chemical shifts with pH). However, the titration curves are not usually reported in the experimental papers; rather, a single pK_a value is given for each group, corresponding to the inflection point. For proteins, the relation between the titration curves and the individual pK_a's is not completely straightforward, as recognized by Tanford, Ullmann, and others.[49,96,92−95] Indeed, coupling between titrating groups can affect the shape of the curves, and one must distinguish carefully between alternate definitions of the proton binding reaction. For example, the binding of a proton to a group of interest can be studied with the other groups held in a fixed protonation state or free to adapt (as here, both in the computations and the cited experiments). The experimental pK_a values are normally the inflection points of the titration curves, so that they correspond to the following "Henderson−Hasselbach" pK_a definition[92,95]

$$pK_{a,i} = pH + \log \frac{x_i}{1 - x_i} \qquad (17)$$

where i represents a titrating group and x_i is the mean population of its protonated form at the given pH. By comparing our computed inflection points with the experimental ones, we are consistent with the experimental definition, and we refer to these inflection points simply as pK_a's.

Our test set includes eight residues for which only experimental upper or lower bounds are available. These residues are Asp54, Asp93, Asp101, and Glu73 in Barnase and Asp7, Asp27, Tyr31, and the C-terminus in OMTKY3. Since precise experimental pK_a's for these residues were not known, they were treated in a special way:[14] an experimental value was assumed to be equal to the computed, if the computed pK_a was in the experimental range; otherwise, it was assumed to be the experimental lower or upper bound. Several theoretical methods were compared to each other and to experiment: a "standard" Poisson approach (denoted SC-PB), using a single protein conformation; residue-GB with a single protein conformation (SC-GB); and our constant-pH, Monte Carlo approach. With this last approach, the pK_a calculations involve three general stages: (1) computing the residue−residue interaction energy matrices; (2) fitting the residue−residue GB interaction energies with a simple function (leading to five matrices of fitting coefficients; see Methods); and (3) the Monte Carlo simulations themselves, which explore the space of side chain conformations and protonation states and yield the average protonation states as a function of pH. This three-stage workflow is schematized in Figure 3. We also consider the Null model and the empirical, PROPKA method.[14]

Table 4 and Figure 4 summarize the agreement between experiment and the various computational methods. Good agreement is obtained with the simplest, Null model, as already shown by many authors.[40,97,32,63] The rms deviation is just 1.07 pH units, and the maximum error is 3.5. The empirical, PROPKA method[14] gives even better results, with an rms error of just 0.88 units, a maximum error of 4.4, and a correlation between PROPKA and experiment of 0.74. With PROPKA, the pK_a shift is expressed as a sum of physicochemical and empirical contributions, which take into account desolvation effects, the number and geometry of hydrogen bonds, and long-range electrostatic interactions. The different terms were fitted to reproduce experimental pK_a's. In fact, three of the six proteins considered here (BPTI, OMTYK3, and HEWL) were also part of the PROPKA parametrization set, which included just two other proteins (RNase A and RNase H). Thus, there may be a favorable bias; however, PROPKA gave results nearly as good, earlier, for proteins not included in the parametrization set.[14]

The "standard" SC-PB method with a protein dielectric value of four gives an rms deviation from experiment of 2.3 pH units, with two large errors: Y53 in 2LZT (error 10.8) and D75 in 1A2P (error 9.7). Excluding these residues, the mean error is 1.8. SC-GB (with the same dielectric) gives a lower deviation of 1.6 units. The maximum error is just 4.3 with SC-GB. The correlation between theory and experiment is also higher with SC-GB, 0.71 compared to 0.67. Increasing the protein dielectric to eight gives significantly lower errors of 1.36 (SC-PB) and 1.30 (3.9), with similar correlations. This is consistent with earlier studies, where a higher protein dielectric gives lower average errors,[40,97] with values as large as 20−80 giving the best results. Notice that this does not necessarily mean that a large protein dielectric is the more physically correct value. Indeed, continuum models contain systematic errors that can make the interpretation less straightforward (see Discussion). Computer simulations and experimental data on dry protein powders indicate that the interior of globular proteins is best represented with a low dielectric value of around 4−8.[10]

The multiconformation GB method gives the best results among the continuum electrostatic approaches considered here. With a protein dielectric of four, the rms error is 1.22, only slightly greater than the Null model; the maximum error is 3.9, slightly lower than PROPKA; and the theory/experiment correlation is 0.77, slightly higher than PROPKA. If the protein dielectric is increased to eight, the quality of the results changes only slightly: the rms error decreases slightly (to 1.16), while the maximum error increases slightly and the correlation decreases slightly. It is satisfactory that good results are obtained with a rather low protein dielectric of four since the side chain degrees of freedom are explicitly represented at the atomic level (through the MC simulations); they should not also be represented implicitly through a high protein dielectric (see Discussion). Finally, the performance is equally good for highly shifted and weakly shifted pK_a's, as reported later on.

For MC-GB, the Null model, and PROPKA, the rms deviations from experiment are fairly close (1.2, 1.1, and 0.9

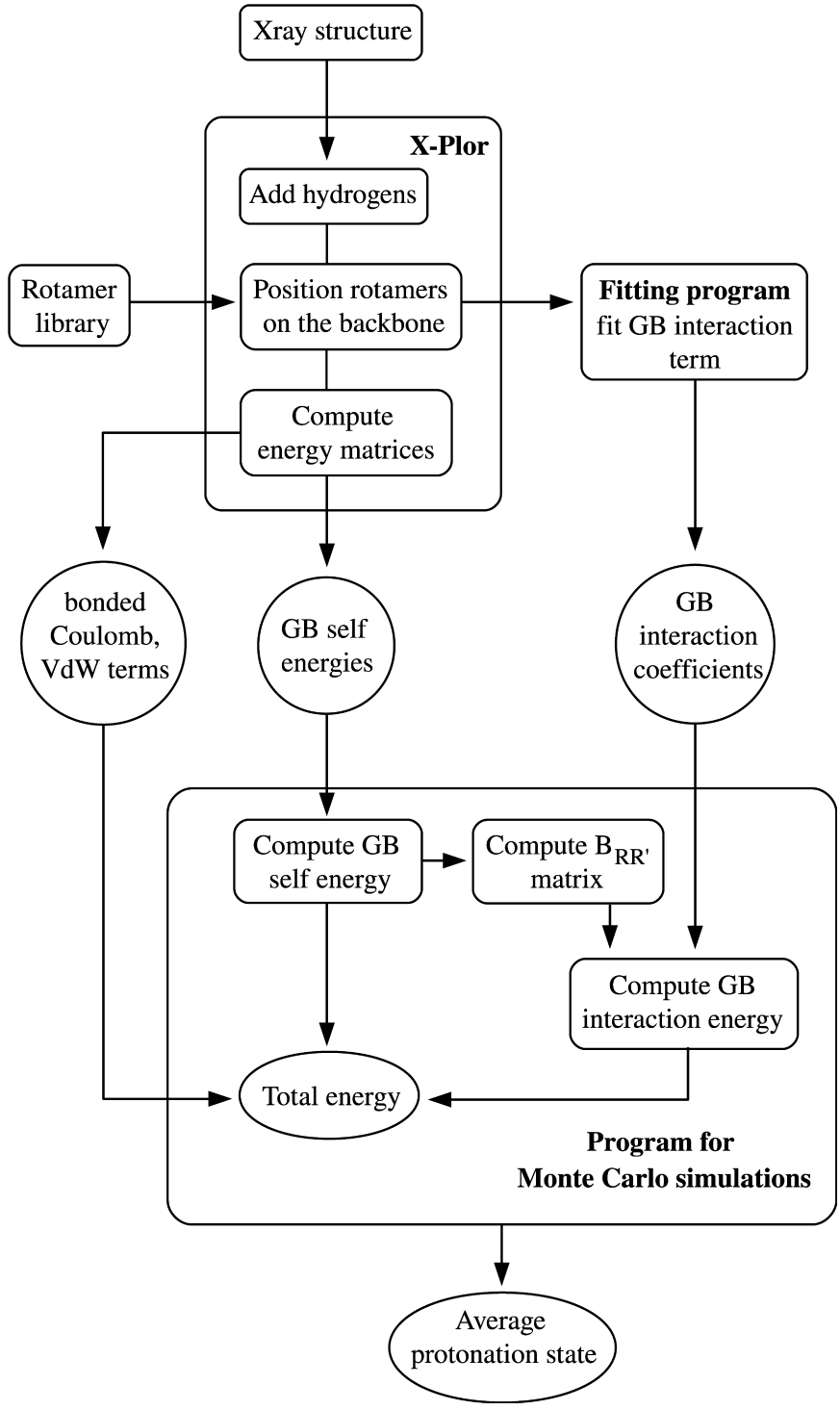

Figure 3. Chart of the computations performed with the multiple-conformation GB method. The input is an X-ray structure from the PDB. The output includes the computed pK_a values for titratable groups.

pK_a units), and so we have done statistical tests to assess the significance of the differences between models. For each model, we considered the square deviations of the 78 computed pK_a's from experiment, say $\{\delta X_i^2\}_{GB}$, $\{\delta X_i^2\}_{Null}$, and $\{\delta X_i^2\}_{PROPKA}$, as random quantities. We used an F-test to establish that their variances are significantly different.[81] Then we used two-tailed Student's tests to test the null hypothesis that the three data sets are all instances of the same random variable δX^2, given the observed differences between their means.[81] The data correspond to about 130 degrees of freedom in the Student's sense, and the t values are 1.56 for GB vs the Null model and 2.24 for GB vs PROPKA. Thus, we may assert with a 99% confidence level that the GB and PROPKA rms deviations are significantly different. Comparing GB and the Null model, the confidence level is only 88%, so that the difference in performance between the two models is only moderately significant.

Table 5 summarizes the results obtained by several other groups for proteins included in our test set. Constant-pH MD simulations by Brooks and co-workers led to better agreement for lysozyme but similar or slightly worse agreement for BPTI, OMTKY3, and barnase.[52,98] Nielsen and Vriend obtained an rms error of about 0.9 for five of our proteins, using a single conformation PB approach with a large protein dielectric of 16.

TABLE 4: Comparison of Different pK_a Methods

method	protein dielectric	rms deviation[a] (maximum)[b]	correlation[c]
single conformation PB	$\varepsilon_p = 4$	2.34 (10.8)	0.67
single conformation GB	$\varepsilon_p = 4$	1.62 (4.3)	0.71
multiconformation GB	$\varepsilon_p = 4$	1.22 (3.9)	0.77
single conformation PB	$\varepsilon_p = 8$	1.36 (5.7)	0.67
single conformation GB	$\varepsilon_p = 8$	1.30 (3.9)	0.70
multiconformation GB	$\varepsilon_p = 8$	1.16 (4.2)	0.71
Null model	–	1.07 (3.5)	–
PROPKA	–	0.88 (4.4)	0.74

[a] rms deviation between computed and experimental pK_a's. [b] In parentheses: maximum error. [c] Pearson correlation coefficient between computed and experimental pK_a's.

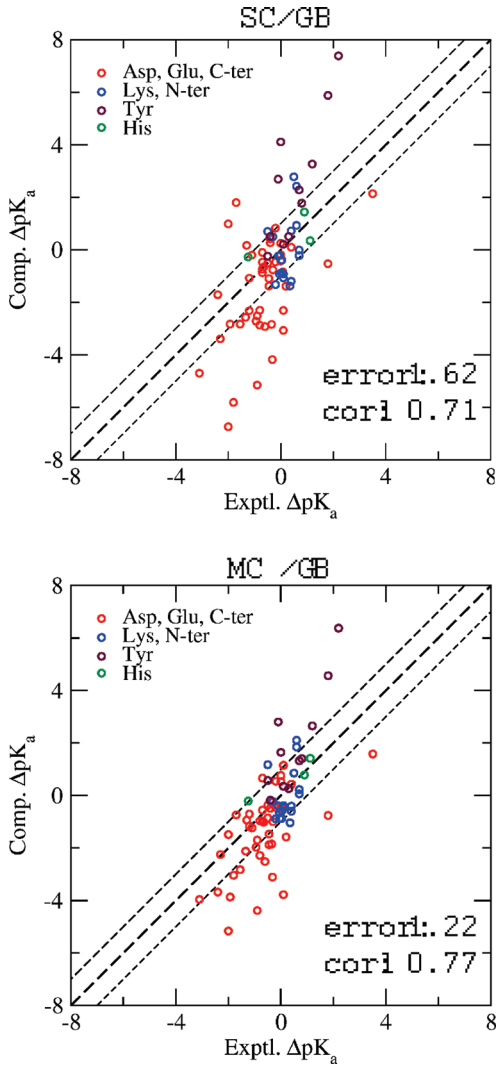

Figure 4. Comparison of the computed and experimental pK_a shifts from the single-conformation method (SC-GB, upper panel) and the multiple-conformation method (MC-GB, bottom panel). The thick dashed line in each inset represents perfect agreement between predictions and experiments; the thin dashed lines are 1 pH unit above and below. The titrating groups are identified by colors. The rms error and the correlation between computed and experimental pK_a shifts are indicated.

A large protein dielectric is expected to give poor results for buried groups and highly shifted pK_a's; these authors' data set included 8 buried groups out of 127, and their rms error was 1.25. The multiconformation PB method of Gunner and coworkers gave an rms error of 0.9 for four proteins.[48] Wisz and Hellinga obtained an rms error of 0.8 for five proteins,[63] using a rather sophisticated, semiempirical method. Warwicker obtained an improved rms error of 0.6 with a hybrid method that treats surface groups with a Debye−Huckel approach and more buried groups with a traditional, SC-PB approach.[99] Finally, very good results were obtained by Spassov and Yan, using a GB solvent and a physiological ionic strength, with an rms deviation of 0.4 for lysozyme, BPTI, and protein G.[100] The use of ionic strength may have helped reduce the error, as noted by Brooks and co-workers.[98] This rapid survey is obviously very incomplete. It does show that improved agreement can be obtained, especially if one uses a higher protein dielectric and/or additional parametrization. Nevertheless, the present MC-GB approach is not much poorer overall, has a good ability to treat highly shifted groups (see below), and can be further improved (see Discussion).

With the MC-GB method, the total CPU time needed for a small protein (1A2P) is about 123 h, including 72 h for the matrix calculations, 12 h for the fitting coefficients, and 39 h for the Monte Carlo simulations. Using a single, eight-core computer, the entire calculation takes under 16 h. Obviously, the empirical models are much faster; for example, the PROPKA calculations take just a few minutes.

We have also analyzed the titration curves obtained with MC-GB and their slopes. The maximum slope of each curve can be interpreted as Hill's coefficient n, which measures the influence of other titrating groups on the group of interest.[5] The titration curves for BPTI are shown in Figure 5. Overall, out of 78 titrating sites, 68 (87%) have n values below 0.85; the rest (13%) have n values between 1.1 and 0.85. In a previous study,[48] the computed (respectively, experimental) slopes had the following distribution: 38% (50%) below 0.85; 60% (48%) between 0.85 and 1.1; and 2% (2%) above 1.1. Thus, our slopes tend to be somewhat too low. For 33 groups with published experimental slopes, our slopes had an rms error of 0.23 and a mean error of −0.13. We found that the computed slopes were sensitive to details of the Monte Carlo procedure, such as the number of conformational relaxation steps following each protonation change. More work is needed to investigate this point.

4.2.2. Large and Small pK_a Shifts. It is important to examine separately groups that have significant pK_a shifts (compared to the usual model compound).[32] Indeed, these are the groups that are hardest to predict and for which the Null model is not satisfactory. In addition, they are often functionally important.[101,49] Out of 78 sites considered in this work, 57 have weak experimental pK_a shifts, $|\Delta pK_a| < 1$ pH unit; 21 have noticeable shifts, $|\Delta pK_a| \geq 1$; and 8 have large shifts, $|\Delta pK_a| \geq 2$. Table 6 summarizes the prediction results.

With SC-PB and a protein dielectric of four, the intermediate group has the largest rms error: 3.0. However, this large value is entirely due to two sites, Y53 in 2LZT and D75 in 1A2P, which give errors of 10.8 and 9.7. For the 20 other sites, the errors are similar to the low and high shift cases. For the highest shifts (eight sites), the predictions are actually a little better than for the low shifts (1.91 vs 2.06). With SC-GB (and a protein dielectric of four), the intermediate group also gives the largest rms error, 1.9, compared to 1.5 for the low and large shifts. With PROPKA, the errors are similar across the whole range of shifts (0.9 units); the largest error is in the low shift group.

Finally, with MC-GB, the performance actually improves slightly as the shifts become larger: the intermediate group has the same rms error as the low shift group (1.2) but a smaller maximum error (2.8), while the larger shifts are reproduced within one pH unit (rms error of 0.93), comparable to PROPKA (but without any empirical parameter adjustment). Thus, by

TABLE 5: RMS Deviation between Experimental and Computed pK_a's: Comparison to Some Earlier Studies[a]

authors	overall	lysozyme	BPTI	OMTKY3	protein G	barnase
Khandogin, Brooks[98b]		−(0.7)	−(0.9)	−(0.6)	−	−(1.0)
Nielsen, Vriend[149]	0.87	0.66	0.60	1.19	0.87	0.90
Georgescu et al.[48]	0.93	0.81	0.67	1.26	0.63	−
Wisz, Hellinga[63]	0.76	0.69	0.45	0.80	0.53	1.17
Warwicker[99]	0.64	0.47	0.35	0.77	0.80	0.76
this work[b]	1.21 (0.86)	1.47 (1.28)	0.77 (0.65)	0.95 (0.52)	0.95 (0.79)	1.55 (0.90)

[a] rms deviations between the computed and experimental pK_a's. [b] Mean unsigned error in parentheses.

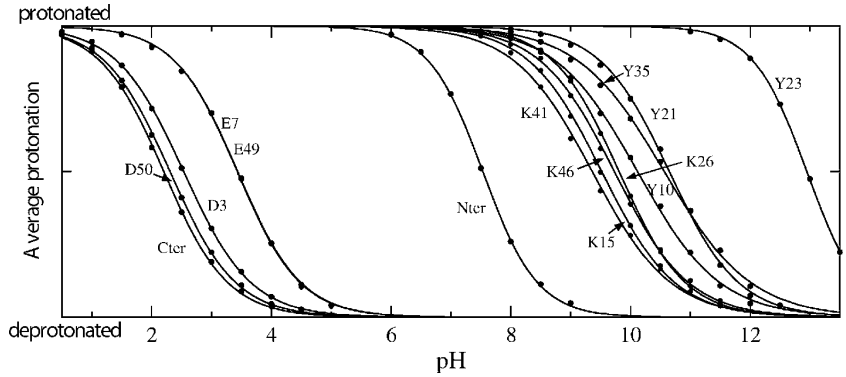

Figure 5. Titration of the 14 amino acids in BPTI. The dots are the populations obtained from the Monte Carlo simulations with MC-GB, and the smooth curves correspond to fits using Hill's expression. Each curve is labeled by the titrating group.

TABLE 6: Performance for Weakly and Strongly Shifted pK_a's[a]

		method			
experimental shift	number of groups	MC-GB	SC-GB	SC-PB	PROPKA
$\|\Delta pK_a\| < 1$	57	1.22 (3.9)	1.52 (4.3)	2.06 (9.7)	0.86 (4.4)
$\|\Delta pK_a\| > 1$	21	1.22 (2.8)	1.85 (4.1)	2.96 (10.8)	0.95 (2.3)
$\|\Delta pK_a\| > 2$	8	0.93 (1.9)	1.47 (3.5)	1.91 (4.4)	0.85 (2.3)
all pK_a's	78	1.22 (3.9)	1.62 (4.3)	2.34 (10.8)	0.88 (4.4)

[a] rms deviations between the computed and experimental pK_a's (maximum deviation in parentheses). Results are shown separately for groups with small and large shifts (relative to the corresponding model compounds).

explicitly representing the conformational reorganization of side chains in response to protonation changes, we improve the prediction quality, as observed by earlier workers.[48,49,65,102,52,103] This is another indication that the MC-GB approach is physically meaningful.

4.2.3. Behavior of Individual Groups. In this section, we describe selected groups and the effect of conformational mobility on their predicted pK_a's. This will help understand the limitations of the current approach and possible ways to improve it.

BPTI Glu49 and Asp50. Using the single conformation GB approach, we compute the pK_a of Glu49 to be 2.1 in moderate agreement with the experimental value of 3.6. Including side chain flexibility gives a pK_a of 3.4, in close agreement with experiment. The structures observed in the Monte Carlo simulations are illustrated in Figure 6. Upon ionization, Glu49 reorients to make a salt bridge with Arg53. The Oδ (Glu49)···NH (Arg53) distance is 4.6 Å at pH = 2 and 3.1 Å at pH = 4, after Glu49 ionization. In the X-ray structure used in this work (PDB entry 4PTI[104]), Glu49 is positioned to interact with its backbone amino group. In the MC simulations, in contrast, Glu49 prefers to be solvent-exposed when protonated or to interact with Arg53 when ionized. Interestingly, in the solution structure (PDB entry 1PIT[105]), determined by NMR, Glu49 also makes a salt bridge with Arg53 in some of the reported conformers.

For Asp50, the MC-GB approach gives a pK_a of 2.3, in good agreement with the experimental value of 3.1; SC-GB gives a pK_a of 1.5, in poorer agreement. In the X-ray structure, Asp50 makes a salt bridge with Arg53 and a hydrogen bond with nearby Ser47, which can explain the experimental pK_a downshift. In the MC simulations, protonated Asp50 does not interact with Arg53 (Oδ(Asp50)···NH_2(Arg53) distance of 5.3 Å), while at a higher pH, deprotonated Asp50 makes a salt bridge with Arg53. The corresponding conformations are illustrated in Figure 6.

Lysozyme Asp66. The experimental pK_a of Asp66 in Lysozyme is 0.9, which corresponds to a very large downshift. The ionized side chain form is stabilized by hydrogen bonds with the hydroxyl groups of Tyr53, Thr69, and Ser60 and with the backbone amino groups of Thr69, Arg68, and Asp66. SC-GB gives a pK_a of −0.7, 1.6 units below the experimental value. With MC-GB, the computed pK_a increases to 0.0, in better agreement with experiment. This improvement is due to a reorientation of the Thr69 hydroxyl upon protonation/deprotonation of Asp66. When Asp66 is protonated, Thr69 donates a hydrogen bond to Gly49. When Asp66 is ionized, Thr69 makes a hydrogen bond to the Asp66 side chain. Since the X-ray structure was determined at a high pH, where Asp66 is always ionized, the positions of the backbone groups are optimized to favor interactions with deprotonated Asp66. With our current implementation of MC-GB, these backbone groups cannot reorganize explicitly upon Asp66 protonation/deprotonation; their reorganization is treated implicitly, through the protein dielectric constant.

This residue was also studied recently by another method that takes into account side chain flexibility (in a more limited way).[102] With a single conformation method, a pK_a of −2.4 was obtained, while the multiconformation method gave a pK_a of 0.9, in perfect agreement with experiment. The improvement was attributed to reorientation of the same, Thr68 hydroxyl group.

Lysozyme Tyr53. The experimental pK_a of this residue is 12.1, upshifted by 1.8 units, compared to the model compound. The

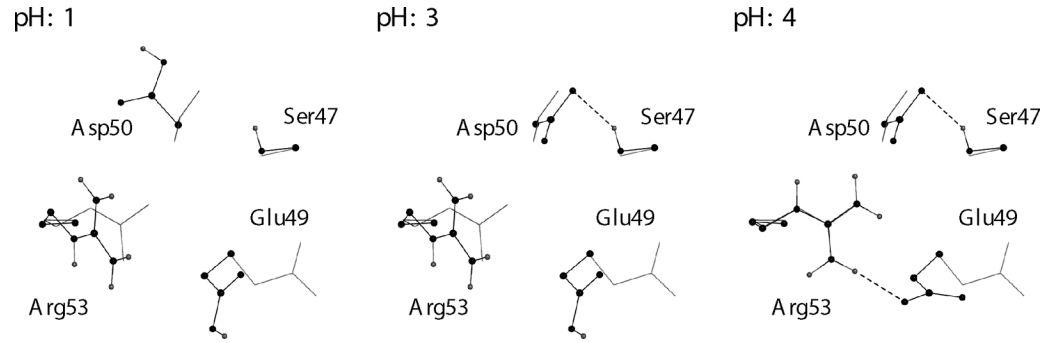

Figure 6. BPTI conformations observed in the Monte Carlo simulations at three selected pH's. The experimental X-ray structure[104] is shown in gray. For clarity, backbone atoms and nonpolar hydrogens are not shown. Polar hydrogens are gray.

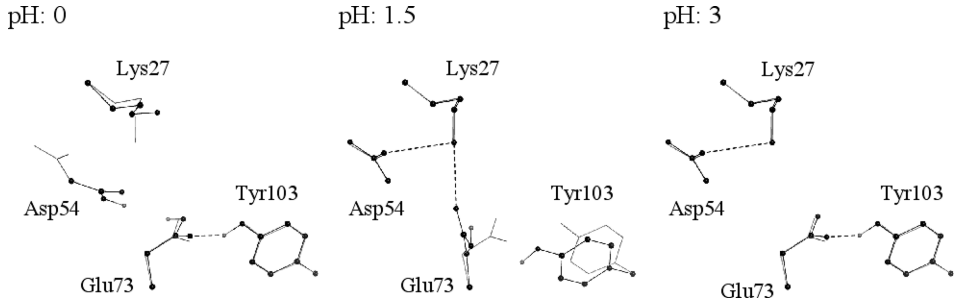

Figure 7. Barnase conformations observed in the Monte Carlo simulations at three selected pH's. The experimental X-ray structure[147] is shown in gray. For clarity, backbone atoms and nonpolar hydrogens are not shown. Polar hydrogens are gray.

upshift can be attributed to the interaction of Tyr53 with Asp66, discussed above. With the experimental structure and SC-GB, we compute the pK_a of this group to be 16.2. Taking into account side chain flexibility, through MC-GB, we obtain 14.9, which is an improvement, but still an overestimate. This could occur because our rotamer library is too limited for tyrosine (only three rotamers differ by more than 30°). Indeed, Tyr53 is partly buried in the X-ray structure, and reorientation toward the solvent may be needed to stabilize the ionized state. In the MC simulations, Tyr53 remained in its crystallographic conformation, possibly because there were not any suitable, more solvent-exposed rotamers in the library. Notice also that at a pH above 12 the protein may begin to unfold, which would then favor Tyr53 ionization. Unfolding is not possible with our rigid-backbone MC approach.

Lysozyme Glu35. The experimental pK_a of this site is 6.2, representing a 1.8 unit upshift. SC-PB with the experimental structure gives a pK_a of 4.6, barely upshifted. With SC-GB and MC-GB, we obtain 3.6 and 3.9, respectively, slightly downshifted. In the X-ray structure, one of the Glu35 side chain oxygens interacts with the Ala10 backbone NH, while the other oxygen is solvent-exposed. Thus, the source of the experimental upshift is not evident.

Barnase Asp54 and Glu73. The experimental pK_a of Asp54 is not known precisely but is less than 2.2. With SC-GB, we predict a very low value of −1.8; with side chain flexibility (MC-GB), we predict a larger value of 1.0, still well within the experimental range. In the X-ray structure (determined at a pH of 7.5, where Asp54 is ionized), Asp54 makes a salt bridge with Lys27 and a hydrogen bond with the backbone NH of the same Lys27. These interactions can explain the downshifted pK_a. In the MC simulations, Asp54 rotates away from Lys27 at low pH, with a distance increasing from 3.0 Å (X-ray structure) to 4.1 Å, and loses its hydrogen bond to the Lys27 backbone; it then makes a weak hydrogen bond to Glu73, with a Oδ(Asp54)···Oε(Glu73) distance of 4.1 Å. This reorganization is illustrated in Figure 7. At pH = 2, where Asp54 becomes ionized, it rotates back to interact with Lys27. The predicted Nζ(Lys27)···Oδ(Asp54) and NH(Lys27)···Oδ(Asp54) distances are 3.3 and 2.9 Å, respectively, close to the X-ray values.

As for Glu73, also shown in Figure 7, we find that it reorients upon protonation, losing a hydrogen bond to Tyr103; this leads to a smaller computed pK_a shift when side chain flexibility is included.

OMTKY3 Asp7 and Asp27. The Asp7 pK_a is not known precisely, but it is less than 2.2. SC-GB predicts 4.2, which is too high. MC-GB predicts 3.1, which is an improvement, though still higher than the experimental upper bound. In the MC simulations at pH = 2, where Asp7 is ionized, one of the Asp7 side chain oxygens interacts with the Ser9 backbone NH; the other one makes a hydrogen bond with the Ser9 side chain hydroxyl.

Another case where SC-GB and MC-GB differ is Asp27, whose experimental pK_a is 2.3. With SC-GB and the X-ray structure, we compute a pK_a of 5.8, so that the protonated form is overstabilized. In the crystal structure, Asp27 interacts with Tyr31, with an Oδ(Asp27)···OH(Tyr31) distance of 2.6 Å. In the MC simulations with deprotonated Asp27, the Tyr31 hydroxyl reorients to donate a hydrogen bond to Asp27. The MC-GB pK_a prediction is 3.3, in much better agreement with experiment.

Thioredoxin Asp20 and Asp26. In oxidized *Escherichia coli* thioredoxin, Asp26 is partly buried and has a large pK_a of 8.1, while solvent-exposed Asp20 has an unshifted pK_a of 4.[106] The titration of these two residues was studied previously by MD free energy simulations in explicit solvent and also by MD simulations with a GB solvent.[33] Here, we obtain an Asp20 pK_a of 3.6, with either SC-GB or MC-GB, in good agreement with experiment. This result also agrees with the previous GB-solvent MD simulations, where the computed pK_a shift was zero[33] (notice that the explicit-solvent MD simulations underestimated the pK_a[33]).

For Asp26, SC-GB predicts a pK_a of 6.1, while MC-GB gives 5.6, slightly farther from the experimental result. This is one of

the largest errors obtained with MC-GB. We do not observe any conformational change when Asp26 becomes protonated, which explains why the SC-GB and MC-GB results are similar. In the earlier MD simulations, when Asp26 became ionized, Lys57 reoriented to form a salt bridge with it; in contrast, protonated Asp26 did not interact closely with Lys (consistent with the X-ray structure, where Asp26 is presumably protonated). This behavior was seen with both explicit and implicit (GB) solvent.[33] Here, in contrast, Lys57 remains 4.6 Å away from ionized Asp26 (Nζ(Lys57)···Oδ(Asp26) distance). Nevertheless, the pK_a shift computed here with MC-GB (+1.6 units) is close to the value obtained from the earlier MD study (pK_a shift of +2.3 units with the GB MD simulations).[33]

5. Discussion

5.1. Hybrid Models: Modeling Dielectric Relaxation. We begin by considering the nature of the hybrid dielectric model used here, with protein backbone, protein electronic, and solvent motions all treated implicitly, while protein side chain motions were treated explicitly. The relation between the explicit and implicit treatments was analyzed in a profound way by Kirkwood and Fröhlich,[107,54,108] and it is instructive to make a brief detour and consider their analysis in the case of a spherical molecule embedded in a uniform solvent.[109−112] The spherical molecule can be thought of as a protein caricature. Using linear response theory, Kirkwood and Fröhlich related the fluctuations of the molecule's dipole moment ΔM, treated explicitly, to the protein and solvent dielectric constants that would be used in an implicit treatment, ε_p and ε_w

$$\frac{\langle \Delta M^2 \rangle}{kTR^3} = \frac{(\varepsilon_p - 1)(1 + 2\varepsilon_w)}{\varepsilon_p + 2\varepsilon_w} \tag{18}$$

Here, R is the protein radius, and the brackets $\langle\rangle$ represent thermal averaging. Fröhlich went further, introducing the first hybrid model, with the fast degrees of freedom treated as a dielectric continuum (dielectric constants ε_p^∞, ε_w^∞ for the protein and solvent) and the slow degrees of freedom treated explicitly. The slow degrees of freedom then obey a relation analogous to eq 18[54,112,113]

$$\frac{\langle \Delta M_{slow}^2 \rangle}{kTR^3} = \frac{(\varepsilon_p - 1)(\varepsilon_p^\infty + 2\varepsilon_w)}{\varepsilon_p + 2\varepsilon_w} - \frac{\varepsilon_w^\infty(\varepsilon_p^\infty - 1)(\varepsilon_p^\infty + 2\varepsilon_w)}{\varepsilon_w(\varepsilon_p^\infty + 2\varepsilon_w^\infty)} \tag{19}$$

In this relation, the separation between explicit (slow) and implicit (fast) degrees of freedom is obtained by a rigorous method, which assumes only that the slow degrees of freedom do not respond to a high frequency electric field.[54] The explicit degrees of freedom are represented by a single quantity, $\langle \Delta M_{slow}^2 \rangle$, which is an average over many atoms and many conformations. Nevertheless, eqs 18 and 19 show how, in a simple case, a thermodynamic property, $\langle \Delta M^2 \rangle$, can be obtained rigorously from a hybrid model.

For protonation free energies, which are of interest here, an even simpler relation exists if the system geometry is simple enough. Specifically, let us view the fast degrees of freedom of protein and solvent as a single, uniform, dielectric continuum.[113,114] This should be a rather accurate approximation for the electronic degrees of freedom of protein and solvent. Indeed, these ideas have been tested since the early days of electron transfer theory.[17,9] In that case, we have the following relation between the free energy ΔG_{MC} computed from the hybrid model and the true free energy, ΔG[113]

$$\Delta G = \Delta G_{MC} + \Delta G^{el} \tag{20}$$

ΔG_{MC} may be obtained (as in Results) from an MC simulation with a uniform dielectric constant ($\varepsilon^\infty = \varepsilon_p^\infty = \varepsilon_w^\infty$), and ΔG^{el} is the free energy when only the fast degrees of freedom are modeled (implicitly, through the continuum model[17,54]). In our calculations and most other pK_a calculations with hybrid models, the ΔG^{el} term can be considered to cancel when the protein and model compound are compared. Thus, with a hybrid model where only the electronic degrees of freedom are treated implicitly, the protonation free energies are obtained from a well-defined approximation.

The next step is to treat *all* the solvent degrees of freedom implicitly, as in most (though not all[115]) hybrid models used to date. This step is actually straightforward since it consists of replacing the potential energy of the protein by a potential of mean force (PMF), as discussed above (Theory).[42] Many studies have shown that a good GB variant is a reasonable approximation to the PMF. Thus, integrating out the solvent degrees of freedom is a well-tested approximation, which can be used in combination with the implicit treatment of the electronic degrees of freedom just discussed.

Finally, we want to integrate out the backbone motions and take them into account through an increased protein dielectric value. While this is a common method, both in pK_a calculations and protein design,[48,67,116,117] the exact nature of the approximation is less clear than the previous two. Compared to the "bath" of valence electrons, it is a bit harder to view the protein backbone as an isotropic, homogeneous medium, uniformly packed throughout the protein interior, with simple, Gaussian fluctuations. Also, any backbone motions displace the side chains as well, so that the dielectric relaxation of the backbone continuum is, in fact, a mixed response that also involves side chain relaxation. Despite its extensive use, there has been little specific testing of this hybrid, "fixed-backbone + continuum" description for its ability to accurately capture the details of a protein's dielectric relaxation. The dynamics of protein backbones have been shown to have a rather harmonic character, if the longer surface loops are excluded,[25] and this is consistent with simple, Gaussian, polarization fluctuations and a linear dielectric response. Furthermore, one early study did analyze the dipolar fluctuations of the backbone of two proteins and compared them to the predictions of dielectric continuum theory, finding a reasonable agreement.[111] The magnitude of the backbone fluctuations was consistent with a dielectric constant of 2−3. In contrast, the fluctuations of ionized protein side chains have a much larger magnitude, consistent with a higher dielectric value of 20 or more.[112] More work is needed to fully test the fixed backbone description; for example, the protein response to perturbing charges could be computed from the hybrid model and compared to detailed, atomistic simulations.[118] For now, we simply view this hybrid treatment as empirically validated by the good agreement with experiment seen for computed pK_a's here and in related studies.[48,49,65,102]

5.2. Limitations of Continuum Models. Continuum models have specific limitations that are worth reviewing briefly, especially since they may or may not be alleviated by the use of a hybrid model. The first limitation is the use of a linear response approximation, which assumes that the electrostatic potential at the proton insertion site has simple, Gaussian

fluctuations.[119,120] This assumption has been tested indirectly for many systems, by comparing computed and experimental pK_a shifts, and it has been tested more directly for a few systems by analyzing electrostatic potentials from MD simulations.[119,33,50,121] Deviations from linear response were observed for a titratable site in thioredoxin[33,122] and for water close to multivalent ions, for example.[123] However, the contribution of a protein's backbone, which is of interest in our hybrid model, has never been analyzed separately. As noted above, flexible loops can have distinctly anharmonic and nonGaussian motions; however, for titratable positions close to such loops, the solvent will often dominate the protonation free energy, so that the backbone behavior is not too important.

A second limitation in most continuum models is the assumption of a homogeneous, isotropic dielectric behavior. Studies of dielectric relaxation in proteins using experiments[124,125] and simulations[110,118,126−130] have provided evidence for spatial inhomogeneity and anisotropy. However, in several cases, these effects could be mostly attributed to the side chains, especially the ionized side chains.[112,118,126,127] The electronic polarizability of the protein was shown to have a rather isotropic and homogeneous behavior,[118] although the early analyses should be repeated with current, superior, polarizability models.[131,132] Another indication that inhomogeneity arises largely from the side chains is the improved pK_a's obtained here and elsewhere[48,49,65,102] when side chain motions are modeled explicitly (or implicitly, but with an improved continuum scheme[99]).

A third, more fundamental limitation is the use of a smoothed, spatially averaged polarization density in the continuum model,[54,133] neglecting the atomic, granular structure of protein and solvent. The effects of granularity have been studied with more detailed, dipole lattice models.[134,135] They are illustrated by the effect of individual, buried waters on pK_a's and by the effect of large, local, conformational rearrangements of the protein.[136,137] Again, by modeling side chain motions explicitly, we expect that local effects can be treated more accurately than by a pure continuum scheme.

Finally, a fourth, serious limitation of continuum models involves the choice of the source charges, which are typically atomic point charges, often taken directly from a molecular mechanics force field. Thus, the smoothing just discussed is applied only to the polarization density and not to the source charges, unlike textbook continuum electrostatics models.[54,133] This choice of charges leads to a consistency problem, which has been analyzed in detail for several systems.[24,138,122,139] Indeed, the charges are usually optimized for use in combination with a low protein dielectric value, ε_p; for example, a molecular mechanics set is optimized for use with $\varepsilon_p = 1$. A low protein dielectric value is then appropriate to describe equilibrium potentials and fields.[139,140] Unfortunately, a low dielectric value is not always appropriate to describe the reorganization that occurs in response to a new charge, such as a titrating proton. Frequently, a larger value of $\varepsilon_p = 4$ or more is required. Thus, in cases where structural reorganization is sizable, continuum calculations using a single protein dielectric value can lead to very large errors for either the equilibrium potentials, the reorganization free energies, or both,[24,138,122,140] resulting in inaccurate pK_a's. This problem has also been analyzed at length by Warshel, Krishtalik, and co-workers, with conclusions that are consistent with ours.[140,32] One way to avoid the problem is to divide the protonation reaction into two distinct steps (following Hush, Marcus, and others), with one step corresponding to proton insertion into a fixed environment and the second step corresponding to relaxation of the environment.[16−19] The two steps can employ different protein dielectric constants, giving good results for several systems.[24,138,122] In the present work, instead, we use a hybrid model, where most of the protein reorganization is described explicitly (through side chain MC exploration); this should alleviate or eliminate the consistency problem.

5.3. Hybrid Model with Residue-GB. This paper has established a method to predict the acid/base behavior of proteins with a hybrid model that treats side chain motions explicitly. Monte Carlo simulations in a semigrand canonical ensemble yield ensembles of structures and protonation states as a function of pH. A large side chain rotamer library is used, allowing an accurate description of side chain rotamer conformational space.

A distinctive feature of our method is the use of a recent, residue-GB variant.[66] An obvious advantage is the increased speed of GB calculations, compared to PB. The initial, atomic-GB variant was parametrized earlier to match a large set of PB calculations.[84] We showed above that the residue-GB variant reproduces PB solvation and conformational free energies at least as well. This good performance is a bit surprising since no reparameterization was done relative to the original, atomic-GB. With residue-GB, the atomic radii are averaged (harmonically) over each residue's side chain or backbone. For the more exposed atoms, with the smaller atomic B's, the residue B will be somewhat increased; for the more buried atoms, the residue B will be somewhat decreased (relative to the atomic B). Earlier workers, using a very similar atomic-GB variant, noticed that the atomic B's were underestimated, compared to an "exact" Poisson calculation; increasing the B values empirically led to better agreement with PB.[141] It may be that the residue B values have the same effect. The more exposed atoms behave better, although the more buried atoms may behave worse; since the exposed atoms make the largest contribution to the solvation energy, the overall effect is an improvement. This remains to be analyzed in more detail.

The residue-pairwise character of the present GB variant allowed us to derive an efficient computational algorithm, inspired by protein design methodology, where matrices of interaction energies and of fitting coefficients are precomputed. As a result, during the Monte Carlo simulations, the fluctuating shape of the protein/solvent boundary is accounted for in a way that is essentially exact, within the GB framework. In particular, there is no need to treat the side chain fluctuations with a mean field approximation; rather, they are sampled rigorously according to the correct, Boltzmann distribution. Obviously, reproducing exactly the GB boundary treatment does not mean that one reproduces the true physical effects of the protein−solvent boundary. Nevertheless, our treatment contrasts with several other recent continuum electrostatic methods that use an effective, average boundary.[48,49,65,102,63,116,117] In fact, until now, the only method that gave an exact treatment of the protein/solvent boundary was constant-pH MD.

As a test set, we computed the pK_a's of 78 residues in 6 proteins for which experimental data are available. The multi-conformation GB approach (MC-GB) gives improved results compared to single-conformation PB and GB: modeling side chain flexibility explicitly improves the pK_a prediction. The good results were obtained using a physically reasonable dielectric constant of four for the protein. This is only slightly larger than the dielectric constant (around 2) estimated from MD simulations for the interior region of several proteins.[112,127,142,143] For these regions, the electronic polarizability and the motions of the backbone polar groups account for most of the dielectric constant. Furthermore, the MC-GB performance does not depend

much on the precise dielectric value, with $\varepsilon_p = 4$ and 8 giving similar results. Importantly, the performance is equally good for pK_a's with large and small shifts. In contrast, for single-conformation PB, the best performance is obtained with a large (and questionable) protein dielectric of 20 or even 80, and the performance for highly shifted pK_a's is poor, as shown by many authors.[40,32]

There are several directions to systematically improve the method, so that further studies are needed. One aspect that should be explored is the use of a larger test set of proteins. Another is to use a still more detailed rotamer library. A third aspect that affects performance is the precise division of the protein atoms into groups, for the calculation of the residue solvation radii. Here, each side chain was treated as a single group, but other choices are possible. Indeed, the residue solvation radius is a harmonic average over the group, so that the more exposed atoms (with the smallest atomic solvation radii) tend to dominate the mean. In the case of a tyrosine, for example, a partly exposed hydroxyl will lead to a small residue solvation radius, even though most of the side chain is buried. This explains why the GB interaction energies are somewhat smaller with residue-GB than with atomic-GB, as shown in the original paper (see Figures 1 and 3 in ref 66). In the future, we will explore the possibility of splitting some of the large side chains into two groups, such as the benzene ring and the hydroxyl group in the tyrosine case.

A significant limitation of the model is the use of a single backbone conformation, with the backbone flexibility modeled implicitly. Using multiple backbone conformations would rapidly increase the computational cost. Indeed, for each backbone conformation, the side chain rotamers have a different meaning, and separate matrices must be computed. A simple implementation would lead to a linear increase with respect to the number N of backbone conformations, both in time and in memory requirements. On the other hand, allowing multiple backbone conformations for limited regions of the protein, such as flexible loops, would be less expensive. Another shortcoming that may be important is the electrostatic model with fixed atomic charges, where electronic polarizability is treated implicitly. More work is needed to determine whether this is a significant source of error. Finally, while nonpolar interactions are included within the protein, we have largely ignored nonpolar effects related to the solvent. In the future, these could be added to the implicit solvent model through surface area terms or other treatments.[42,84]

Overall, the method has a clear physical basis, and there are several directions to systematically improve it. In its present form, it already gives reasonable accuracy for pK_a's and should be a useful tool to help increase our understanding of protein electrostatics.

Acknowledgment. G.A. and S.P. acknowledge financial support through a PENEK reinforcement 0505/04 grant "Modification of the specificity of synthetase family proteins with biomolecular simulations". G.A. and T.S. acknowledge support through Cyprus-France ZENON bilateral cooperation funds (project "Protein design simulations with application to the redesign of the genetic code").

Supporting Information Available: Additional methods and results, and additional tables and figures. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Sorensen, S. P. L.; Hoyrup, M.; Hempel, J.; Palitzsch, S. *C. R. Trav. Lab. Carls.* **1917**, *12*, 68–163.
(2) Linderstrøm-Lang, K. *C. R. Trav. Lab. Carls.* **1924**, *15*, 1–2.
(3) Pace, C. N.; Grimsley, G. R.; Scholtz, J. M. *J. Biol. Chem.* **2009**, *284*, 13285–13289.
(4) Lehninger, A.; Cox, M.; Nelson, D. L. *Principles of Biochemistry*; Freeman: New York, 2008.
(5) Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; Freeman: New York, 1999.
(6) Perutz, M. *Mechanisms of cooperativity and allosteric regulation in proteins*; Cambridge University Press: Cambridge, 1990.
(7) Kyte, J. *Structure in protein chemistry*; Garland Publishing: New York, 1995.
(8) Varadarajan, R.; Zewert, T. E.; Gray, H. B.; Boxer, S. G. *Science* **1989**, *243*, 69–72.
(9) Bendall, D. S., Ed. *Protein electron transfer*; BIOS Scientific Publishers, 1996.
(10) Simonson, T. *Rep. Prog. Phys.* **2003**, *66*, 737–787.
(11) Marcus, R. *Annu. Rev. Phys. Chem.* **1964**, *15*, 155–196.
(12) Warshel, A.; Parson, W. *Q. Rev. Biophys.* **2001**, *34*, 563–679.
(13) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. *Science* **2004**, *303*, 186–195.
(14) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.
(15) Huang, R. B.; Du, Q. S.; Wang, C. H.; Liao, S. M.; Chou, K. C. *Protein Eng.* **2010**, *23*, 35–42.
(16) Hush, N. *Trans. Faraday Soc.* **1961**, *57*, 557–580.
(17) Marcus, R. *J. Chem. Phys.* **1956**, *24*, 979–989.
(18) Warshel, A. *J. Phys. Chem.* **1982**, *86*, 2218–2224.
(19) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430–437.
(20) Tanford, C.; Kirkwood, J. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.
(21) Warshel, A.; Russell, S. *Q. Rev. Biophys.* **1984**, *17*, 283–342.
(22) Schaefer, M.; Froemmel, C. *J. Mol. Biol.* **1990**, *216*, 1045–1066.
(23) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.
(24) Simonson, T.; Archontis, G.; Karplus, M. *J. Phys. Chem. B* **1999**, *103*, 6142–6156.
(25) Brooks, C. L.; Karplus, M.; Pettitt, M. *Adv. Chem. Phys.* **1987**, *71*, 1–259.
(26) Becker, O., Mackerell, A., Jr., Roux, B., Watanabe, M., Eds. *Computational Biochemistry & Biophysics*; Marcel Dekker: New York, 2001.
(27) Warshel, A. *Computer modelling of chemical reactions in enzymes and solutions*; John Wiley: New York, 1991.
(28) Field, M. J. *A Practical Introduction to the Simulation of Molecular Systems*; Cambridge University Press: New York, 2007.
(29) Mulholland, A. *Drug Discovery Today* **2005**, *10*, 1393–13402.
(30) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368–8372.
(31) Sham, Y.; Chu, Z.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458–4472.
(32) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44*, 400–417.
(33) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.
(34) Ghosh, N.; Cui, Q. *J. Phys. Chem. B* **2008**, *112*, 8387–8397.
(35) Zheng, L.; Chen, M.; Yang, W. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20227–20232.
(36) Ji, C. G.; Mei, Y.; Zhang, J. Z. H. *Biophys. J.* **2008**, *95*, 1080–1088.
(37) Delepierre, M.; Dobson, C.; Karplus, M.; Poulsen, F.; States, D.; Wedin, R. *J. Mol. Biol.* **1987**, *197*, 111–130.
(38) Havranek, J.; Harbury, P. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11145–11150.
(39) Warwicker, J.; Watson, H. *J. Mol. Biol.* **1982**, *157*, 671–679.
(40) Antosiewicz, J.; McCammon, J.; Gilson, M. *J. Mol. Biol.* **1994**, *238*, 415–436.
(41) Schaefer, M.; Vlijmen, H. W. T. v.; Karplus, M. *Adv. Protein Chem.* **1998**, *51*, 1–57.
(42) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
(43) Gilson, M.; Davis, M.; Luty, B.; McCammon, J. *J. Phys. Chem.* **1993**, *97*, 3591–3600.
(44) Cramer, C.; Truhlar, D. *Chem. Rev.* **1999**, *99*, 2161–2200.
(45) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.
(46) You, T.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721–1733.
(47) Beroza, P.; Case, D. A. *J. Phys. Chem.* **1996**, *100*, 20156–20163.
(48) Georgescu, E. R.; Alexov, E.; Gunner, M. *Biophys. J.* **2002**, *83*, 1731–1748.
(49) Kim, J.; Mao, J.; Gunner, M. *J. Mol. Biol.* **2005**, *348*, 1283–1298.
(50) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523–544.
(51) Börjesson, U.; Hünenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.
(52) Lee, M.; Salsbury, F., Jr.; Brooks, C., III *Proteins* **2004**, *56*, 738–752.

(53) Mongan, J.; Case, D. A.; McCammon, J. *J. Comput. Chem.* **2004**, *25*, 2038–2048.
(54) Fröhlich, H. *Theory of Dielectrics*; Clarendon Press: Oxford, 1949.
(55) Roux, B.; Beglov, D.; Im, W. *Simulation and theory of electrostatic interactions in solution*; Pratt, L., Hummer, G., Eds.; American Institute of Physics, 1999; pp 473−491.
(56) David, L.; Luo, R.; Gilson, M. *J. Comput. Chem.* **2000**, *21*, 295–309.
(57) Still, W. C.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
(58) Hawkins, G. D.; Cramer, C.; Truhlar, D. *Chem. Phys. Lett.* **1995**, *246*, 122–129.
(59) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
(60) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
(61) Bashford, D.; Case, D. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
(62) Feig, M.; Brooks, C. L., III *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
(63) Wisz, M. S.; Hellinga, H. *Proteins* **2003**, *51*, 360–377.
(64) Zollars, E. S.; Marshall, S. A.; Mayo, S. L. *Protein Sci.* **2006**, *15*, 2014–2018.
(65) Barth, P.; Alber, T.; Harbury, P. B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4898–4903.
(66) Archontis, G.; Simonson, T. *J. Phys. Chem. B* **2005**, *109*, 22667–22673.
(67) Dahiyat, B. I.; Mayo, S. L. *Science* **1997**, *278*, 82–87.
(68) Resat, H.; Mezei, M. *J. Am. Chem. Soc.* **1994**, *116*, 7451–7452.
(69) Woo, H. J.; Dinner, A.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392–6400.
(70) Collins, M. D.; Hummer, G.; Quillin, M. L.; Matthews, B. W.; Gruner, S. M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16668–16671.
(71) Hill, T. *Introduction to Statistical Thermodynamics*; Addison-Wesley: Reading, Massachussetts, 1962.
(72) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.
(73) Simonson, T. *J. Phys. Chem. B* **2000**, *104*, 6509–6513.
(74) Allen, M.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon Press: Oxford, 1991.
(75) Frenkel, D.; Smit, B. *Understanding molecular simulation*; Academic Press: New York, 1996.
(76) Sham, Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744–1753.
(77) Cornell, W.; Cieplak, P.; Bayly, C.; Gould, I.; Merz, K.; Ferguson, D.; Spellmeyer, D.; Fox, T.; Caldwell, J.; Kollman, P. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(78) Guérois, R., Lopez de la Paz, M., Eds. *Protein Design: Methods And Applications*; Humana Press: Totowa, NJ, 2007.
(79) Schmidt am Busch, M.; Lopes, A.; Mignon, D.; Simonson, T. *J. Comput. Chem.* **2008**, *29*, 1092–1102.
(80) Brünger, A. T. *X-plor version 3.1, A System for X-ray crystallography and NMR*; Yale University Press: New Haven, 1992.
(81) Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986.
(82) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. *J. Biomol. Struct. Dyn.* **1991**, *8*, 1267.
(83) Moulinier, L.; Case, D. A.; Simonson, T. *Acta Cryst. D* **2003**, *59*, 2094–2103.
(84) Lopes, A.; Aleksandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. *Proteins* **2007**, *67*, 853–867.
(85) Swanson, J.; Adcock, S.; McCammon, J. *J. Chem. Theory Comput.* **2005**, *1*, 484–493.
(86) Wagner, F.; Simonson, T. *J. Comput. Chem.* **1999**, *20*, 322–335.
(87) Brünger, A. T.; Karplus, M. *Proteins* **1988**, *4*, 148–156.
(88) Bashford, D. *Scientific Computing in Object-Oriented Parallel Environments, volume 1343 of Lecture Notes in Computer Science*; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; Springer: Berlin, 1997; pp 233−240.
(89) Bashford, D. *Front. Biosci.* **2004**, *9*, 1082–1099.
(90) Beroza, P.; Case, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804–5808.
(91) Madura, J.; Briggs, J.; Wade, R.; Davis, M.; Luty, B.; Ilin, A.; Antosiewicz, J.; Gilson, M.; Baheri, B.; Scott, L.; McCammon, J. *Comput. Phys. Commun.* **1995**, *91*, 57–95.
(92) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192–2198.
(93) Onufriev, A.; Case, D. A.; Ullmann, M. *Biochemistry* **2001**, *40*, 3413–3419.
(94) Ullmann, M. *J. Phys. Chem. B* **2003**, *107*, 1263–1271.
(95) Bombarda, E.; Ullmann, M. *J. Phys. Chem. B* **2010**, *114*, 1994–2003.
(96) Sondergaard, C. R.; McIntosh, L. P.; Pollastri, G.; Nielsen, J. E. *J. Mol. Biol.* **2008**, *376*, 269–287.
(97) Warwicker, J. *Protein Sci.* **1999**, *8*, 418–425.
(98) Khandogin, J.; Brooks, C. L. *Biochemistry* **2006**, *45*, 9363–9373.
(99) Warwicker, J. *Protein Sci.* **2004**, *13*, 2793–2805.
(100) Spassov, V.; Yan, L. *Protein Sci.* **2008**, *17*, 1955–1970.
(101) Harris, T. K.; Turner, G. J. *IUBMB Life* **2002**, *53*, 85–98.
(102) Kieseritzky, G.; Knapp, E. W. *Proteins* **2008**, *71*, 1335–1348.
(103) Machuqueiro, M.; Baptista, A. M. *Proteins* **2008**, *72*, 289–298.
(104) Marquart, M.; Walter, J.; Deisenhofer, J.; Bode, W.; Huber, R. *Acta Crystallogr.* **1983**, *B39*, 480.
(105) Berndt, K. D.; Guntert, P.; Orbons, L. P.; Wüthrich, K. *J. Mol. Biol.* **1992**, *227*, 757–775.
(106) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. *Proteins* **2002**, *48*, 388–403.
(107) Kirkwood, J. *J. Chem. Phys.* **1939**, *7*, 911–919.
(108) Simonson, T. *Int. J. Quantum Chem.* **1999**, *73*, 45–57.
(109) Simonson, T.; Perahia, D.; Brünger, A. T. *Biophys. J.* **1991**, *59*, 670–90.
(110) King, G.; Lee, F.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366–4377.
(111) Smith, P.; Brunne, R.; Mark, A.; van Gunsteren, W. *J. Phys. Chem.* **1993**, *97*, 2009–2014.
(112) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1082–1086.
(113) Lleontyev, I. V.; Stuchebrukov, A. A. *J. Chem. Phys.* **2009**, *130*, 085102.
(114) Leontyev, I. V.; Rostov, M. V. V.; Basilevsky, M. V.; Newton, M. D. *J. Chem. Phys.* **2003**, *119*, 8024.
(115) Xin, W. D.; Juffer, A. H. *J. Comput. Phys.* **2007**, *223*, 416–435.
(116) Vizcarra, C. L.; Mayo, S. L. *Curr. Opin. Chem. Biol.* **2005**, *9*, 622–626.
(117) Vizcarra, C. L.; Zhang, N. G.; Marshall, S. A.; Wingreen, N. S.; Zeng, C.; Mayo, S. L. *J. Comput. Chem.* **2008**, *29*, 1153–1162.
(118) Simonson, T.; Perahia, D. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.
(119) Simonson, T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6544–6549.
(120) Song, X.; Chandler, D.; Marcus, R. A. *J. Phys. Chem.* **1996**, *100*, 11954–11959.
(121) Nilsson, L.; Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13867–13872.
(122) Archontis, G.; Simonson, T. *Biophys. J.* **2005**, *88*, 3888–3904.
(123) Sandberg, L.; Edholm, O. *J. Chem. Phys.* **2002**, *116*, 2936–2944.
(124) Hass, M.; Jensen, M.; Led, J. *Proteins* **2010**, *48*, 6482–6494.
(125) Hernandez, G.; Anderson, J. S.; LeMaster, D. M. *Biochemistry* **2009**, *48*, 6482–6494.
(126) Simonson, T.; Perahia, D. *Faraday Discuss.* **1996**, *103*, 71–90.
(127) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118*, 8452–8458.
(128) Voges, D.; Karshikoff, A. *J. Chem. Phys.* **1998**, *108*, 2219–2227.
(129) Song, X. *J. Chem. Phys.* **2002**, *116*, 9359–9363.
(130) Golosov, A.; Karplus, M. *J. Phys. Chem. B* **2007**, *111*, 1482–1490.
(131) Ponder, J.; Case, D. A. *Adv. Protein Chem.* **2003**, *66*, 27.
(132) Friesner, R. A. *Adv. Protein Chem.* **2006**, *72*, 79–104.
(133) Jackson, J. *Classical electrodynamics*; Wiley: New York, 1975.
(134) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211–217.
(135) Sandberg, L.; Edholm, O. *Biophys. J.* **2010**, *98*, 470–477.
(136) Dwyer, J. J.; Gittis, A. G.; Karp, D. A.; Lattman, E. E.; Spencer, D. S.; Stites, W. E.; Garcia-Moreno, B. E. *Biophys. J.* **2000**, *79*, 1610–1620.
(137) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; Garcia-Moreno, B. *Biophys. J.* **2007**, *92*, 2041–2053.
(138) Archontis, G.; Simonson, T. *J. Am. Chem. Soc.* **2001**, *123*, 11047–11056.
(139) Simonson, T. *Photosynth. Res.* **2008**, *97*, 21–32.
(140) Krishtalik, L.; Kuznetsov, A.; Mertz, E. *Proteins* **1997**, *28*, 174–182.
(141) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.
(142) Pitera, J.; Falta, M.; van Gunsteren, W. *Biophys. J.* **2001**, *80*, 2546–2555.
(143) Park, H.; Jeon, Y. H. *Phys. Rev. E* **2007**, *75*, 021916.
(144) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721–4729.
(145) Lee, T. W.; Qasim, M. A.; Laskowski, M.; James, M. N. *J. Mol. Biol.* **2007**, *367*, 527–546.
(146) Ramanadham, M.; Sieker, L. C.; Jensen, L. H. *Acta Crystallogr.* **1990**, *B46*, 63–69.
(147) Martin, C.; Richard, V.; Salem, M.; Hartley, R.; Mauguen, Y. *Acta Crystallogr.* **1999**, *D55*, 386–398.
(148) Katti, S.; Lemaster, D.; Eklund, H. *J. Mol. Biol.* **1990**, *212*, 167.
(149) Nielsen, J. E.; Vriend, G. *Proteins* **2001**, *43*, 403–412.

JP104406X