JCTC Journal of Chemical Theory and Computation

Benchmarking pK_a Prediction Methods for Residues in Proteins

Courtney L. Stanton and Kendall N. Houk*

Department of Chemistry and Biochemistry, University of California Los Angeles, 607 Charles E. Young Drive East, Los Angeles, California 90095

Received January 2, 2008

Abstract: Methods for estimation of pK_a values of residues in proteins were tested on a set of benchmark proteins with experimentally known pK_a values. The benchmark set includes 80 different residues (20 each for Asp, Glu, Lys, and His), half of which consists of significantly variant cases (Δp$K_a \geq 1$ pK_a unit from the amino acid in solution). The method introduced by Case and co-workers [*J. Am. Chem. Soc.* **2004**, *126*, 4167−4180], referred to as the molecular dynamics/generalized-Born/thermodynamic integration (MD/GB/TI) technique, gives a root-mean-square deviation (rmsd) of 1.4 pK_a units on the benchmark set. The use of explicit waters in the immediate region surrounding the residue was shown to generally reduce high errors for this method. Longer simulation time was also shown to increase the accuracy of this method. The empirical approach developed by Jensen and co-workers [*Proteins* **2005**, *61*, 704−721], PROPKA, also gives an overall rmsd of 1.4 pK_a units and is more or less accurate based on residue type—the method does very well for Lys and Glu, but less so for Asp and His. Likewise, the absolute deviation is quite similar for the two methods—5.2 for PROPKA and 5.1 for MD/GB/TI. A comparison of these results with several prediction methods from the literature is presented. The error in pK_a prediction is analyzed as a function of variation of the pK_a from that in water and the solvent accessible surface area (SASA) of the residue. A case study of the catalytic lysine residue in 2-deoxyribose-5-phosphate aldolase (DERA) is also presented.

I. Introduction

Ionizable residues play a critical role in many of the important physical and chemical properties of proteins including folding and stability,[1–3] protein−protein interactions,[4] substrate binding,[5] and enzymatic reaction mechanisms.[6] Consequently, accurate pK_a prediction methods are of great interest for understanding pH-dependent properties of proteins and in the fields of rational drug and protein design.[7]

The pK_a value of an ionizable group can vary significantly from its value in solution due to the altered environment of the interior of the protein. These variant cases are not only the most difficult to predict, but are often the most interesting. One example can be found in the enzyme 2-deoxyribose-5-phosphate aldolase (DERA), which catalyzes the reaction shown in Figure 1.[8,9] The first step of the reaction involves nucleophilic attack by unprotonated Lys167. Lysine in solution is protonated at neutral pH, with a pK_a of 10.5. This value is perturbed to around 7 in the active site of the enzyme, allowing the reaction to occur. The environment of Lys167 is quite complicated, making it difficult to predict the pK_a.

The free energy profile of proton binding is dominated by electrostatic contributions from intraprotein interactions and protein−solvent interactions.[10] Explicit treatment of electrostatic interactions for every pair of charges in a fully atomistic model of both protein and solvent is computationally very expensive even with a classical force-field and was indeed completely infeasible before recent advances in computer power and electrostatic treatments such as the particle mesh Ewald procedure.[11] Therefore, most of the current developments in pK_a prediction have focused on implicit electrostatic treatments, especially solutions to the

* Corresponding author. E-mail: houk@chem.ucla.edu.

10.1021/ct8000014 CCC: $40.75
Published on Web 05/09/2008

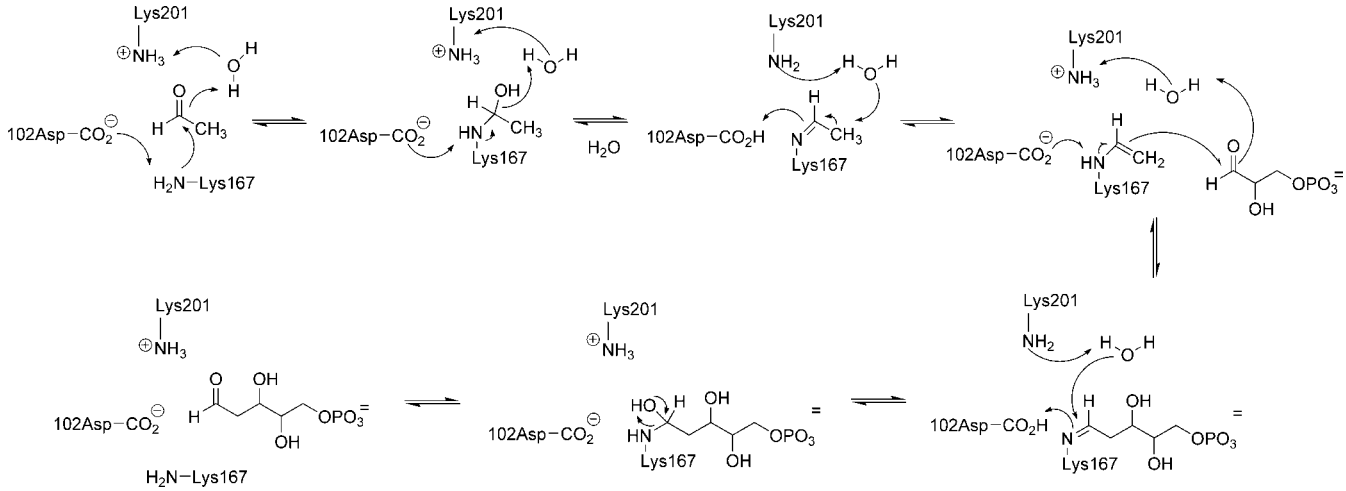

Figure 1. Mechanism of 2-deoxyribose-5-phosphate aldolase as proposed by Heine et al.[8,9]

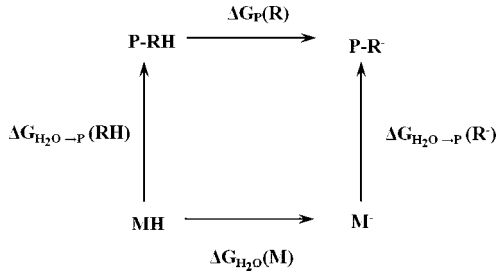

Figure 2. Thermodynamic cycle used to calculate pK_a shifts. MH refers to the model compound in aqueous solution, P-RH refers to the residue in the protein environment. p$K_{a,R}$ is the pK_a value for the residue in the protein, and p$K_{a,M}$ is the pK_a of the model compound in aqueous solution. The change in pK_a is calculated as shown in eq 1.

Poisson−Boltzmann equation (PBE),[12–16] which will be discussed in further detail in the Theoretical Background section.

This study was designed to test and compare recent promising pK_a prediction methods for the four ionizable residues Asp, Glu, Lys, and His. We present calculations from the literature and new calculations using two different methods: one method introduced by Case and co-workers[17] based on molecular dynamics and thermodynamic integration and another introduced by Jensen and co-workers,[18] which is purely empirical. The benchmark set includes equal numbers of two different groups of residues: (1) residues that have an experimental pK_a that does not vary significantly from the amino acid in solution (actually a model compound in solution; see theTheoretical Background section) with a ΔpK_a of <1, referred to as low variants, and (2) residues that vary significantly with a ΔpK_a of ≥ 1, referred to as high variants. A brief theoretical summary of pK_a prediction of residues in proteins is presented, followed by a more detailed description of the various methods.

II. Theoretical Background

Generally, pK_a prediction methods are based on the thermodynamic cycle shown in Figure 2.[19] The pK_a value of a residue in a protein (p$K_{a,R}$) is calculated relative to the experimentally determined pK_a of a model compound in aqueous solution (p$K_{a,M}$). The model compound is typically the amino acid side chain with neutral blocking groups meant to account for the backbone substituent effect in proteins, which decreases the pK_a of titratable residues[20] (see Figure 3 in the Computational Methods section). According to the cycle, the relative pK_a can be determined by calculating either (1) the difference in the free energy change of proton loss from the residue in the protein, $\Delta G_P(R)$, compared to proton loss of the model compound in solution, $\Delta G_{H_2O}(M)$, or (2) the difference in the free energy change of the protonated residue being transferred from the aqueous environment to the protein environment, $\Delta G_{H_2O} \rightarrow P(RH)$ compared to the unprotonated residue being transferred from the solvent to the protein, $\Delta G_{H_2O} \rightarrow P(R^-)$. Both strategies have been used to predict pK_a values.

$$\Delta \mathrm{p}K_a = (\mathrm{p}K_{a,R} - \mathrm{p}K_{a,M}) = \frac{1}{2.3}RT\,[\Delta G_P(R) - \Delta G_{H_2O}(M)] = \frac{1}{2.3}RT\,[\Delta G_{H_2O\rightarrow P}(R^-) - \Delta G_{H_2O\rightarrow P}(RH)] \quad (1)$$

Most studies report a root-mean-square deviation (rmsd) from experiment of ≤ 1 pK_a unit. However, this is somewhat misleading considering that most of the data are dominated by residues that are on the surface of the protein or residues that do not have strong neighboring intramolecular interactions and, therefore, do not generally vary significantly from the pK_a value of the model compound in solution. Residues that do vary substantially from p$K_{a,M}$ typically have more complex interactions and are therefore more difficult to model accurately.

As mentioned in the Introduction, solving the electrostatics for a fully atomistic model of a macromolecule such as a protein in solution is computationally quite expensive, due to the long-range nature of electrostatic interactions. Most methods have focused on decreasing this cost by introducing approximations to the full electrostatic treatment. The free energy changes in eq 1 are generally treated as purely electrostatic and can be broken down and evaluated in a number of ways. For example, Demchuk and Wade[21] describe the free energy change of transferring the protonated (or unprotonated) residue from the aqueous to the protein environment in this way:

$$\Delta\Delta G\,(\mathrm{MH} \rightarrow \mathrm{P\text{-}RH}) = \Delta\Delta G_{\mathrm{Born}} + \Delta\Delta G_{\mathrm{Boltzmann}} + \Delta\Delta G_{\mathrm{dipole}} + \Delta\Delta G_{\mathrm{charge}} \quad (2)$$

where ΔG_{Born} is the free energy to transfer a charge from deionized solvent to a neutral cavity with no permanent dipoles, $\Delta G_{\mathrm{Boltzmann}}$ represents the charge−charge interaction with the solvent−ions, $\Delta G_{\mathrm{dipole}}$ is the charge−charge interaction with the permanent dipoles, and $\Delta G_{\mathrm{charge}}$ is the charge−charge interaction in the protein.

In order to accurately calculate these free energies, one must account for a number of effects in both the solvent and in the protein, including induced dipoles,[22,23] structural relaxation,[24] Debye−Hückel screening,[25] and hydrogen bonding.[26] There are two ways to deal with these effects, either implicitly, treating them as an average macroscopic property, or explicitly, treating them as a microscopic property. The most popular methods follow a hybrid approach by approximating some or all of these effects implicitly and some explicitly. Many of these methods solve the linearized Poisson−Boltzmann equation (LPBE)[27,28] using numeric finite difference techniques.[29,30] In this framework, the protein is modeled as a low dielectric cavity with an assigned "protein dielectric" constant and is surrounded by a high dielectric medium, such as water, in which the distribution of counterions is described by a Boltzmann distribution. The electrostatic potential of the protein and solvent are calculated and the interaction energy is obtained by assigning fixed atomic charges in the protein and calculating the interaction with the protein and solvent potentials. These calculations are typically done using a molecular mechanics force-field, while the boundary between protein and solvent is determined by the atomic coordinates of the protein or model compound.

The Poisson−Boltzmann (PB) methods vary in which effects are modeled explicitly and which are approximated by adjusting the protein dielectric parameter. Assignment of the protein dielectric remains controversial, and values between 2 and 80 have been reported.[10,28,31–33] Presumably, a large protein dielectric can account for the protein relaxation and screening of electrostatic interactions;[24,34–36] this works well for solvent exposed residues and residues without significant protein charge−charge interactions, but less so for residues with more complex interactions. Theoretically, the more microscopic detail included in the model, the smaller the protein dielectric should be. If all interactions are treated explicitly, a protein dielectric equal to 1 (vacuum) should be used.

Demchuk and Wade[21] evaluated the effect of varying the protein dielectric and found that a homogeneous dielectric is not sufficient to account for the loss of microscopic detail for all residues and that the appropriate dielectric constant depends on the extent of solvent exposure. Recent methods have begun to account for the heterogeneity of the protein reaction field by introducing varying degrees of microscopic detail. Some methods retain the macroscopic continuum approach but modify the potentials used to calculate the interaction energy. Other methods account for structural reorganization by incorporating conformational sampling techniques such as molecular dynamics or Monte Carlo sampling. There are also examples of fully atomistic models that treat all the interactions explicitly. Specific examples are discussed in the Description of Methods section.

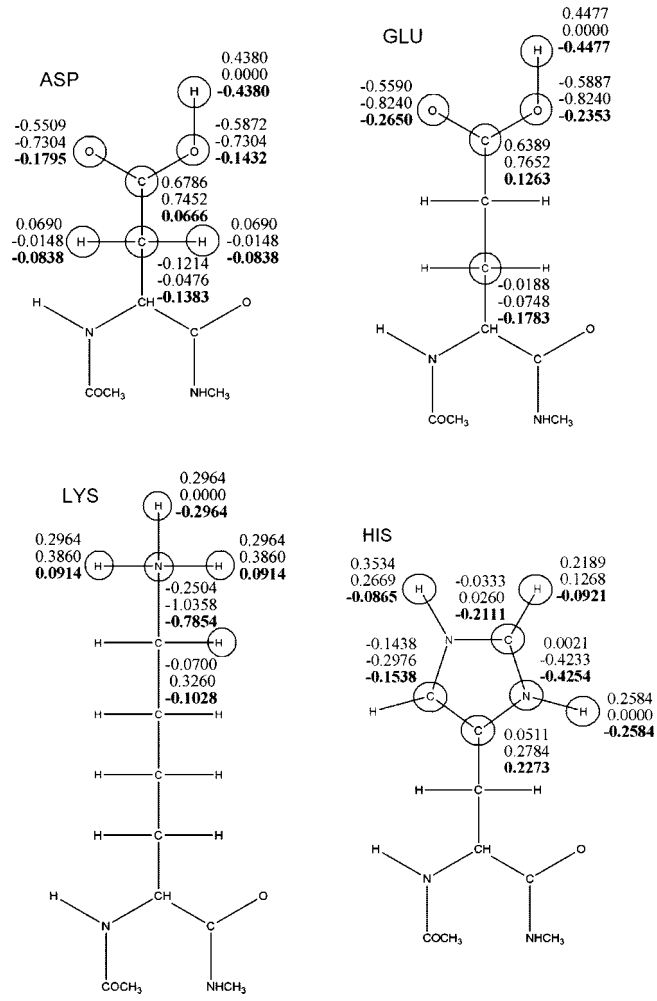

Figure 3. Model systems used for Asp, Glu, Lys, and His. The circled atoms were perturbed in the thermodynamics integration calculations for both the model system in water and the side chain in protein. The top number is the partial charge of the atom in the protonated state, the middle number is the charge in the unprotonated state, and the bottom number is the difference between the two (with some modifications to make sure that the total change was −1). Atoms were perturbed if the charge difference between the charged and uncharged state was >0.05.

The importance of choosing an appropriate benchmark set has been previously discussed.[37] If a benchmark study is dominated by surface residues (typically low variants), any model that employs a high protein dielectric constant will appear to give accurate results, since the pK_a of the model compound is also modeled in a high dielectric medium, such as water. For example, if a benchmark set includes a majority of low variants, even a model that predicts every Asp residue to have a pK_a of 4 (ΔpK_a = 0) will appear to be a good model regardless of whether it models more complex interactions properly. Presumably, it is important for a benchmark to include a significant number of high variants in order to test the ability of the model to predict unscreened charge−charge interactions.

III. Description of Methods

This section provides a description of some of the recent most promising approaches to calculating pK_a values. The methods that are later compared in the Results and Discussion section are introduced, as well as several other methods of interest.

A. Methods Addressing Protein Dielectric Heterogeneity. Mehler and Guarnieri developed a potential that analyzes the unique microenvironment surrounding each residue.[38] Specifically, they characterize the hydrophilicity or hydrophobicity around each titratable residue and use this information to modify the electrostatic potential at each site. Their method is based on sigmoidally screened Coulomb potentials (SCP) and requires less computational effort than solving the full LPBE. An rmsd of 0.5 pK_a units is reported for the method for a benchmark set of 103 experimental values in seven proteins.

Wisz and Hellinga addressed the issue of dielectric heterogeneity by introducing geometry-dependent dielectric constants for each pairwise interaction.[39] They also fit their model to an extensive set of experimental values to determine empirical parameters that take into account local structure. The benchmark set includes 260 ionizable residues in 41 different protein crystal structures. An rmsd of 0.95 pK_a units is reported.

B. Methods Addressing Conformational Sampling. Simonson and co-workers recently used a combined PB/linear response approximation (LRA) approach using molecular dynamics simulations to account for structural relaxation.[40] They predict the pK_a value for three different residues, two of which are high variants, using a small protein dielectric of 1 and 2. The average value from two calculations (with the dielectric equal to 1 or 2) gives an rmsd of 1.5 pK_a units. Pokala and Handel introduced the EGAD program as a method for protein design, and tested their electrostatic calculations against a benchmark set of experimental pK_a values.[41] A generalized-Born (GB) continuum model was used,[42] which is a much faster approximation to the PBE. This method is combined with a self-consistent mean field (SCMF) approach for rotamer optimization to account for side-chain relaxation. The reported rmsd is 0.92 pK_a units for 200 ionizable groups from 15 proteins. Gunner and co-workers published a method which combines Monte Carlo sampling with continuum electrostatics using a protein dielectric of 4 to give a reported rmsd of 0.83 pK_a units for 166 residues in 12 proteins.[43]

C. Methods Avoiding Protein Dielectric. Other methods attempt to avoid the errors incurred by using the protein dielectric as an approximation to the dielectric response. Merz used molecular dynamics/free energy perturbation simulations to calculate pK_a values with somewhat limited success, giving an rmsd of 2.8 pK_a units for 2 residues.[44] Molina and co-workers recently developed an accurate quantum mechanics/molecular mechanics (QM/MM) method, which calculates the reaction path of proton loss using quantum mechanics to model the active residue while the rest of the system is treated with a molecular mechanics force-field.[45] They report an rmsd of 0.3 pK_a units for five experimental values. This is also the most expensive method that we have discussed due to the inclusion of quantum mechanical electronic structure calculations. The development of QM/MM techniques using free energy perturbation to predict pK_a values has been pursued by others, including Cui and co-workers.[46] Most of these have met limited success or have not been tested on a benchmark set of significant size.

D. PROPKA and MD/GB/TI, Two Extremes in Computational Prediction of pK_a's. Two methods were tested here for our complete benchmark set, and these were also compared to values computed in the literature by other methods.

PROPKA. The PROPKA method is an empirical approach to calculating pK_a values developed by Jensen and co-workers.[18] It involves standard parameters for adjustments to the pK_a by residues in the vicinity of the ionizable group. The method is extremely fast and has a reported rmsd of 0.89 pK_a units. The benchmark set used to test the method includes experimental values for 314 residues in 44 proteins. The pK_a value is calculated by adding "environmental perturbations" to the pK_a value of the model residue in solution (p$K_{a,model}$) (these perturbations are referred to as ∂pK_a, not to be confused with the ΔpK_a that has been discussed previously as the change in pK_a from solution to protein):

$$\mathrm{p}K_a = pK_{a,\mathrm{model}} + \partial\mathrm{p}K_a \tag{3}$$

where ∂pK_a is the sum of the individual perturbations. Perturbations are calculated for three environmental factors: hydrogen bonding, desolvation effects, and charge−charge interactions. Each hydrogen bond is assigned a perturbation value (∂pK_a) described by a simple distance/angle function multiplied by an empirically determined parameter. The perturbation value for desolvation effects is determined by assessing how many protein atoms are within a given distance of the ionizable residue and multiplying by a parameter. Charge−charge interactions between buried pairs of residues are incorporated in a similar way—a perturbation value is assigned for each charged residue within a given distance and multiplied by an empirically determined parameter. The final pK_a is calculated by adding all of the perturbations to the p$K_{a,model}$ value. The parameters for each perturbation type were optimized empirically. This method is especially attractive for high throughput applications such as protein design.

MD/GB/TI. Case and co-workers recently published a method of pK_a prediction for protein residues using molecular dynamics free energy calculations to simulate a fully atomistic description of the entire protein.[17] While the protein electrostatics and other nonbonded forces are explicitly modeled in all of the simulations with a molecular mechanics force-field, the solvent is modeled either explicitly, as a periodic water box using the TIP3P water model,[47] or implicitly, using the generalized-Born (GB) continuum water model.[48] The GB model is an approximation to solving the PB equation. Free energies are calculated with the thermodynamic integration (TI) technique (see the Computational Methods section for a description of TI). This method will be referred to as the molecular dynamics/generalized-Born/thermodynamic integration (MD/GB/TI) technique. The authors report an rmsd of approximately 1 pK_a unit using

the MD/GB/TI method for three aspartic acid residues, two of which were previously shown to be very difficult to predict with other methods, including the LRA study discussed above in the Methods Addressing Conformational Sampling section.[40] Here, we explore the performance of this method more generally.

IV. Computational Methods

A. Preparation of Protein Atomic Coordinates. All atomic coordinates were downloaded from the PDB database, and the files were manually stripped of any solvent molecules, cofactors, metal ions, or inhibitors. The PDB identifiers and protein names in our benchmark set are the following, with corresponding references to the experimental pK_a determination: 4LZT hen egg-white lysozyme,[49,50] 2RN2 bacterial RNase H,[51,52] 1PPF turkey ovomucoid inhibitor,[53,54] 1BEO fungal beta-cryptogein,[55] 1PGA bacterial protein G B1 domain,[56] 3RN3 bovine RNase A,[20,57] 1DE3 fungal RNase alpha-sarcin,[58] 2TRX bacterial thioredoxin (oxidized),[59] 1A2P bacterial barnase,[60] 1ANS sea anemone neurotoxin III,[61] 1RGA fungal RNase T1,[62,63] 1HNG rat CD2,[64] 1XNB bacterial xylanase,[65] 2SNM bacterial nuclease mutant,[66] 1BTL bacterial beta-lactamase,[67] 1MUT bacterial MutT,[68] 1NFN human Apo E3,[69] 1FEZ bacterial phosphonoacetaldehyde hydrolase,[70] 1GS9 human Apo E4,[71] 1LE2 human Apo E2,[70] 1NZP human DNA polymerase lambda lyase domain,[72] 2BCA bovine calbindin D9K,[73] 2EBX snake erabutoxin b,[74] 3SSI bacterial proteinase inhibitor Ssi,[75] 1STN bacterial nuclease,[76] 1ERT human thioredoxin (reduced),[77] 1DG9 bovine PTPase,[78] 1L54 phage T4 lysozyme mutant,[79] 2LZM phage T4 lysozyme.[80]

B. PROPKA Calculations. PROPKA is a freely accessible program provided by the Jensen group at the University of Copenhagen.[18] Here, we used the web-based version PROPKA1.0.1.[81] Atomic coordinates can be retrieved from the PDB database by entering the PDB code or uploaded manually in the PDB format directly from the browser. Structures are automatically stripped of all nonprotein molecules, and the program calculates the pK_a for every ionizable residue (Asp, Glu, His, Lys, Tyr, Arg) in the protein within seconds. Here, the PDB files were edited manually to remove all nonprotein atoms, and, in the case of NMR structures or structures with more than one conformation for any residue, an average structure was submitted to the site. The program output reports the pK_a values and the various environmental perturbations used in the calculation.

C. MD/GB/TI Calculations. *Protein Calculations.* Protein crystal structures were downloaded from the PDB database, and the initial protonation states for each of the ionizable residues were assigned before running the molecular dynamics (MD) simulations. PROPKA, which predicts the pK_a values of Asp, Glu, Lys, His, Tyr, and Arg, was used for this purpose (generally speaking Tyr and Arg do not change protonation states in proteins; Asp, Glu, and Lys change protonation states rarely; and His is found quite frequently in either protonation state due to its "normal" pK_a of 6.3). The ionization state was determined from the PROPKA output: Asp, Glu, or His with predicted pK_a > 6.8 were protonated while Lys, Tyr, or Arg with values <7 were unprotonated in the MD simulations.

The MD simulations were performed with AMBER 8.0,[82] which includes a thermodynamic integration utility. Addition of some explicit waters improved results in certain cases (see the Results and Discussion section). Waters were added to the crystal structures using the AMBER module XLeap. The solvatecap command was used to add waters (model WAT) to within an 8 Å radius around the given residue with a van der Waals closeness parameter of 0.4 Å. This generally resulted in addition of 10−20 water molecules.

The OBC version (named for the authors: Onufriev, Bashford, and Case)[83] of the GB implicit solvent model was used. Two GB radii were tested—mbondi and mbondi2 as defined in the AMBER 8.0 literature.[82] Two AMBER forcefields were tested—ff99 and ff03. On the basis of the results of preliminary tests, ff03 and mbondi2 were used for the calculations reported here.

The thermodynamic integration utility was used to calculate the free energy change on going from the protonated state of the residue in question ($\lambda = 0$) to the unprotonated state ($\lambda = 1$). The integral describing the free energy change shown in eq 4 can be solved numerically as shown in eq 5. Dynamics were performed for three values of λ (0.11270, 0.5000, 0.88729). Equation 5 was solved with the corresponding weights (ω) of 0.27777, 0.44444, 0.27777, respectively, for each λ value. These values of λ and the weights have been determined to be optimal for calculating the free energy difference in a thermodynamic integration scheme.[84] Equation 6 describes the dependence of the potential function on λ, where V_0 is the Hamiltonian in the original protonated state and V_1 is the Hamilitonian in the unprotonated state and is referred to as linear mixing.[82]

$$\Delta G = G(\lambda = 1) - G(\lambda = 0) = \int_0^1 \left\langle \frac{\partial V}{\partial \lambda} \right\rangle \tag{4}$$

$$\Delta G \approx \sum_{i=1}^{n} \omega_i \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_i \tag{5}$$

$$V(\lambda) = (1 - \lambda)V_0 + \lambda V_1 \tag{6}$$

Equilibration dynamics were run for 200 ps, and production dynamics were run for an additional 200 ps. No restraints were used and the proteins were free to fluctuate. All protein structures were examined at the end of the simulations for structural integrity, which was maintained in all cases when the GB solvent model was used. There were rare instances when the protein structure denatured with the addition of some internal explicit water molecules, and these cases were not used in the benchmark set.

The change in protonation state is represented as a total change in charge of −1 (change in the van der Waals contribution from the disappearing proton is ignored). The change in charge is confined to the residue in question by altering the partial charges of specific atoms in the residue. Figure 3 shows the partial charges for each residue type for both the protonated and unprotonated states (the figure shows the model compound, but the same charges were used for both the model compound in solution and the residue in the protein).

Model Calculations. The model compounds are shown in Figure 3 and include the titratable amino acid with —CONHCH$_3$ and —NHCOCH$_3$ blocking groups. The change in free energy required to deprotonate the model compound in solution (ΔG_{H_2O}(M) in eq 1) for each residue type is determined using the same MD/GB/TI methodology as described above for protein residues. The value for ΔG_{H_2O}(M) for each residue type is calculated once and used as a reference to calculate the ΔpK_a for each protein residue in the benchmark set, using the following intrinsic pK_a values for each residue type (p$K_{a,M}$ in eq 1): Asp 4.0, Glu 4.4, Lys 10.5, and His 6.3.[28]

SASA Calculations. The percent of solvent accessible surface area (SASA) for each residue in the benchmark set was calculated with the program GETAREA 1.1.[85] GETAREA is a freely accessible web-based program provided by the Sealy Center for Structural Biology at the University of Texas.[86] Protein atomic coordinates must be supplied in PDB format and uploaded to the site for calculation. The same PDB files that were used for the PROPKA calculations were uploaded to the GETAREA Web site, and the SASA per residue was calculated simultaneously for every residue in the protein. The percent SASA is reported as the ratio of the side-chain surface area to the "random coil" surface area for that residue type. The random coil value of residue type X is defined as the average surface area of X in the tripeptide Gly-X-Gly in an ensemble of 30 random conformations. The values are listed in the GETAREA manual.[87]

V. Results and Discussion

A. Guide to Table 1. The calculated ΔpK_a values for 80 residues in 30 different proteins are presented in Table 1. The table is split into 4 subtables for residue types Asp, Glu, Lys, and His. For each type, there are 10 residues that are experimentally known to be high variants ($\Delta pK_a \geq 1$), and 10 residues that are low variants ($\Delta pK_a < 1$). The high variants are shown in bold face font. The data from columns labeled "MD/GB/TI w/waters", "MD/GB/TI w/out waters", and PROPKA were obtained in this work. The other columns represent values taken from the literature. Columns labeled "err" are the difference between predicted and experimental ΔpK_a values. The root-mean-square deviation (rmsd), mean absolute deviation (MAD), and maximum absolute deviation (MAX) from the experiment are shown in red in the bottom three rows and are calculated first according to residue type, and also as a total for all residues in blue at the end of the four tables. The reported values are ΔpK_a values, where $\Delta pK_a = pK_{a,R} - pK_{a,M}$. The MD/GB/TI and PROPKA methods were tested on the entire benchmark set. The other methods listed in Table 1 did not have data available for every residue in the benchmark set used here. The number of values used to calculate the rmsd and MAD are noted in parentheses. The first column is color-coded to show the extent of solvent accessibility. The percent SASA for each residue was calculated as described in Computational Methods: purple indicates <20% accessibility, green is >50% accessibility, and blue is ≤50% or ≥20%.

B. Methods. *MD/GB/TI.* The total rmsd of 1.4 pK_a units for the MD/GB/TI method using only the GB solvent model is relatively high compared to the PB methods. The GB approximation accounts for the solvent in a continuum way and does not always accurately model microscopic interactions such as hydrogen bonding. The worst predictions (noted with an asterisk in Table 1), with an error ≥1.5 pK_a units, were repeated with some explicit waters around the residue in question, as described in Computational Methods. The new values where the explicit water calculations were performed are entered in the column labeled "MD/GB/TI w/waters" in Table 1, but the entire benchmark set was not rerun with explicit waters due to the additional computational cost. Inclusion of explicit waters generally improved the predictions. The rmsd value for the 21 residues that were repeated was reduced from 2.4 to 1.9 pK_a units, without and with waters, respectively. The maximum absolute deviation was also significantly reduced from 5.1 to 2.5. In several cases, inclusion of explicit waters did not change the prediction significantly. When the predictions did change significantly, they were generally improved: 1DE3 Glu96, 1FEZ Lys53, 1NZP Lys312, 1STN His121, 4LZT His15, 3RN3 His48, 1DE3 His104. Exceptions to this include 4LZT Glu7 and 1DE3 His137, which were predicted somewhat less accurately with explicit waters. Inclusion of some explicit waters in the interior of the protein was used previously with a PB method.[88] The results from that study showed that experimental pK_a values could be predicted with a smaller protein dielectric when several explicit waters were used, indicating a more accurate microscopic model.

Figure 4 shows a plot of the data for the MD/GB/TI method without explicit waters. The experimental $\Delta pK_{a,exp}$ are plotted against the predicted $\Delta pK_{a,predict}$ values. The least-squares line through the origin is shown in solid black. The dashed line has a slope of 1 and is shown for comparison.[89] The R^2 value is 0.48, indicating that the two variables share 48% of their variability in common. Most of the data points fall above the dashed line in both of the plots. This indicates that the MD/GB/TI method generally overestimates the pK_a regardless of residue type. For example, Lys66 of PDB entry 2SNM was predicted to have a change in pK_a of −3.1, while the experimental value is −4.1. This means the actual pK_a value is predicted to be 7.4, which is higher than the experimental value of 6.4. The same is true for the majority of the MD/GB/TI predictions and may be the consequence of a systematic error in the method.

The method followed here involved a 200 ps equilibration and 200 ps production simulation for each value of λ. The issue of convergence was explored for several residues by extending the production run to 1 ns for each value of λ. The residues were the following: 1A2P Glu60, 1L54 Lys102, 4LZT His15, 3RN3 Asp14, and 3RN3 His48. For two of the cases, Glu60 and Asp14, the calculated ΔpK_a did not change significantly (<0.5 pK_a units). The other three predictions, Lys102, His15, and His48 were significantly improved by 1, 1.6, and 1.2 pK_a units, respectively. This indicates that many of the runs may not be fully converged and longer simulations may further reduce the error associated with this method. The entire benchmark set was not

Table 1. Comparison of Experimental pK_a Values with Several Different Prediction Methods

PDB Code (residue)	Exp. ΔpK_a	MD/GB/TI w/ waters (This work, Simonson et al.) ΔpK_a		MD/GB/TI w/out waters* (This work, Simonson et al.) ΔpK_a		PROPKA (This work, Jensen et al.) ΔpK_a		Geom dep dielectric (Hellinga et al.) ΔpK_a		Microenv SCP (Guarnieri et al.) ΔpK_a		EGAD (Handel et al.) ΔpK_a		MCCE (Gunner et al.) ΔpK_a		QM/MM (Molina et al.) ΔpK_a	
ASP		calc	err[a]	calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	err
3RN3 (asp14)*	-2.2	-1.0	1.2	-0.3	1.9	-2.6	-0.4	-1.6	0.4	-1.4	0.6			-3.4	-1.2		
4LZT (asp87)	-1.9			-0.8	1.1	-1.5	0.4	-1.1	0.9	-0.9	1.1	-1.1	0.8	-2.8	-0.9		
1PPF (asp27)*	-1.8	0.2	2.0	0.6	2.4	-1.6	0.2	-0.6	1.2	-0.1	1.7	-1.0	0.8	-0.7	1.1	-2.1	-0.3
1XNB (asp11)	-1.5			-0.1	1.4	-2.0	-0.5	-0.8	0.7			-0.4	1.1				
1BEO (asp21)	-1.5			-1.1	0.4	-2.6	-1.1							1.1	2.6	-1.5	0.0
4LZT (asp18)	-1.3			-0.8	0.5	-1.2	0.1	-0.2	1.1	-0.5	0.8	-0.5	0.8	-1.0	0.3		
1XNB (asp106)	-1.3			-1.2	0.1	-1.0	0.3	-0.5	0.8			-0.5	0.8				
1PGA (asp22)	-1.1			-0.4	0.7	-1.8	-0.7	0.2	1.3	-1.1	0.0	-0.5	0.6	-1.8	-0.7		
3RN3 (asp121)*	-0.9	1.0	1.9	1.0	1.9	-0.3	0.6	-1.8	-0.9	-0.1	0.8			-0.8	0.1		
1A2P (asp75)	-0.9			0.3	1.2	-5.3	-4.4	-1.6	-0.7			2.3	3.2	0.5	1.4		
2RN2 (asp94)*	-0.8			0.3	1.1	-1.3	-0.5	-1.3	-0.5	-0.5	0.3			-0.2	0.6		
1PGA (asp47)	-0.6			-0.1	0.5	-1.3	-0.7	-0.1	0.5	-1.4	-0.8	-0.7	-0.1	-1.7	-1.1		
3RN3 (asp53)	-0.3			1.0	1.3	-0.6	-0.3	-0.5	-0.2	0.0	0.3			-0.3	0.0		
4LZT (asp52)[c]*	-0.3	2.1	2.4	2.0	2.3	-1.0	-0.7	0.8	1.1	-0.5	-0.2	-0.4	-0.1	-0.1	0.2		
1PGA (asp36)	-0.2			0.6	0.8	-0.1	0.1	-0.1	0.1	0.4	0.6	0.5	0.7	1.0	1.2		
2TRX (asp20)	-0.2			0.0	0.2	-1.5	-1.3										
1DE3 (asp59)	0.1			-0.1	-0.2	-0.8	-0.9	0.8	0.7			-1.6	-1.7				
1DE3 (asp57)	0.3			-0.1	-0.4	0.0	-0.3	-0.5	-0.8			-0.6	-0.9				
2RN2 (asp10)	2.1			2.7	0.6	3.0	0.9			1.9	-0.2			6.4	4.3		
2TRX (asp26)	4.1			3.7	-0.4	1.2	-2.9	1.9	-2.2								
RMSD (N)[b]		1.9 (4)		1.2 (20)		1.3 (20)		1.0 (17)		0.8 (12)		1.2 (13)		1.6 (14)		0.2 (2)	
MAD		1.9		1.0		0.9		0.8		0.6		1.0		1.0		0.3	
MAX		2.4		2.4		4.4		2.2		1.7		3.2		4.3		0.3	

Table 1. Continued

PDB Code (residue)	Exp. ΔpK_a	MD/GB/TI w/ waters		MD/GB/TI w/out waters		PROPKA		Geom dep dielectric		Microenv SCP		EGAD		MCCE		QM/MM	
GLU		calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	Err
3RN3 (glu2)	-1.8			-2.0	-0.2	-1.7	0.1	-0.5	1.3	-0.6	1.2			-3.1	-1.3		
4LZT (glu7)*	-1.5	0.6	2.1	0.0	1.5	-0.7	0.8	-1.1	0.4	-0.9	0.6	-1.8	-0.3	-0.9	0.6	-1.7	-0.2
1PPF (glu19)	-1.2			-0.3	0.9	-0.5	0.7	-0.2	1.0	-0.3	0.9	-0.7	0.5	-2.8	-1.6	-1.7	-0.5
2RN2 (glu57)	-1.2			0.2	1.4	-1.8	-0.6	0.6	1.8	-1.7	-0.5			-1.9	-0.7		
1A2P (glu60)*	-1.2	0.4	1.6	0.7	1.9	-0.6	0.6	-1.3	-0.1			-1.2	0.0	-2.6	-1.4		
3RN3 (glu111)	-0.9			0.5	1.4	0.2	1.1	-0.4	0.5	0.0	0.9			-0.5	0.4		
2RN2 (glu129)	-0.8			-0.5	0.3	-0.9	-0.1	-1.0	-0.2	-1.4	-0.6			-1.6	-0.8		
2RN2 (glu61)	-0.5			0.5	1.0	-0.8	-0.3	-0.9	-0.4	-0.8	-0.3			-1.5	-1.0		
3RN3 (glu9)*	-0.4	1.6	2.0	1.3	1.7	0.2	0.6	-0.6	-0.2	0.2	0.6			1.0	1.4		
1BCA (glu26)	-0.3			0.4	0.7	0.4	0.7							-1.7	-1.4		
1PPF (glu10)	-0.3			0.7	1.0	0.0	0.3	-0.6	-0.3	-0.1	0.2	-0.3	0.0	-0.9	-0.6	-0.1	0.2
2RN2 (glu119)*	-0.3	1.3	1.6	1.6	1.9	-0.9	-0.6	-1.0	-0.7	-0.6	-0.3			-1.3	-1.0		
1PGA (glu27)	0.1			1.3	1.2	-1.2	-1.3	-0.7	-0.8	-1.3	-1.4	-0.1	-0.2	-0.6	-0.7		
1PPF (glu43)	0.4			0.3	-0.1	0.1	-0.3	-0.1	-0.5	0.0	-0.4	1.0	0.6	0.1	-0.3	0.1	-0.3
1DE3 (glu96)[c]*	0.7	0.7	0.0	2.4	1.7	0.6	-0.1	-0.1	-0.8			-0.3	-1.0				
1ANS (glu20)	1.0			0.7	-0.3	0.1	-0.9	-0.1	-1.1								
1RGA (glu28)	1.5			1.2	-0.3	-0.3	-1.8	-0.1	-1.6			1.3	-0.2				
4LZT (glu35)[c]	1.8			1.8	0.0	0.6	-1.2	0.8	-1.0	1.9	0.1	1.8	0.0	1.8	0.0		
1HNG (glu41)	2.3			2.3	0.0	0.3	-2.0					-1.0	-3.3	1.4	-0.9		
1XNB (glu172)[c]	2.3			2.7	0.4	2.9	0.6	0.3	-2.0			3.2	0.9				
RMSD (N)		1.6 (5)		1.1 (20)		0.9 (20)		1.0 (18)		0.7 (13)		1.1 (11)		1.0 (16)		0.3 (4)	
MAD		1.5		0.9		0.7		0.8		0.6		0.6		0.9		0.3	
MAX		2.1		1.9		2.0		2.0		1.4		3.3		1.6		0.5	

Table 1. Continued

PDB Code (residue)	Exp. ΔpK$_a$	MD/GB/TI w/ waters		MD/GB/TI w/out waters		PROPKA		Geom dep dielectric		Microenv SCP		EGAD		MCCE		QM/MM	
LYS		calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	Err
2SNM (lys66)	-4.1			-3.1	1.0	-2.6	1.5	-2.6	1.5								
1L54 (lys102)*	-3.9	-1.4	2.5	-0.9	3.0	-2.6	1.3	-2.1	1.8								
1MUT (lys39)[c]*	-2.1	0.4	2.5	0.3	2.4	0.0	2.1										
1NFN (lys146)	-1.3			-0.8	0.5	0.0	1.3							-1.1	0.2		
1FEZ (lys53)[c]*	-1.2	0.2	1.4	1.8	3.0	-2.4	-1.2										
1GS9 (lys146)	-1.1			-0.8	0.3	0.0	1.1										
1LE2 (lys143)	-1.1			-0.2	0.9	-0.5	0.6										
1NFN (lys143)	-1.0			-0.3	0.7	0.0	1.0							-2.4	-1.4		
1NZP (lys312)[c]*	-1.0	0.0	1.0	0.7	1.7	-0.2	0.8										
1GS9 (lys143)	-0.6			-0.3	0.3	0.0	0.6										
1LE2 (lys146)	-0.6			-0.5	0.1	-0.1	0.5										
1PPF (lys34)	-0.4			-0.2	0.2	-0.3	0.1			0.5	0.9			-3.3	-2.9		
4LZT (lys33)	-0.1			-0.4	-0.3	-0.2	-0.1	-0.1	0.0	0.9	1.0			-0.7	-0.6		
2BCA (lys41)	0.3			0.2	-0.1	-0.1	-0.4	0.5	0.2	0.0	-0.3			0.1	-0.2		
4LZT (lys96)	0.3			0.1	-0.2	-0.3	-0.6	0.1	-0.2	0.2	-0.1			0.8	0.5		
1PGA (lys28)	0.4			0.2	-0.2	-0.6	-1.0	0.7	0.3	0.9	0.5			1.2	0.8		
2BCA (lys16)	0.4			0.7	0.3	-0.6	-1.0	1.0	0.6	0.2	-0.2			0.7	0.3		
1PPF (lys55)	0.6			0.2	-0.4	0.0	-0.6			0.0	-0.6			-0.3	-0.9		
2BCA (lys7)	0.7			0.9	0.2	0.0	-0.7	0.8	0.1	0.4	-0.3			0.4	-0.3		
2BCA (lys55)	1.3			1.3	0.0	0.0	-1.3	0.9	-0.4	0.7	-0.6			1.2	-0.1		
RMSD (N)		2.0 (4)		1.2 (20)		1.0 (20)		0.8 (10)		0.6 (9)		(0)		1.1 (11)		(0)	
MAD		1.9		0.8		0.9		0.6		0.5				0.7			
MAX		2.5		3.0		2.1		1.8		1.0				2.9			

Table 1. Continued

PDB Code (residue)	Exp. ΔpK_a	MD/GB/TI w/ waters		MD/GB/TI w/out waters		PROPKA		Geom dep dielectric		Microenv SCP		EGAD		MCCE		QM/MM	
HIS		calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	err	calc	Err
3EBX (his6)*	**-3.5**			-4.6	-1.1	0.0	3.5	-0.3	3.2								
3SSI (his43)	**-3.1**			-3.1	0.0	-1.4	1.7	-0.6	2.5								
1STN (his121)*	**-1.0**	0.9	1.9	1.4	2.4	-1.0	0.0	1.2	2.2			1.8	2.8				
4LZT (his15)*	-0.9	1.3	2.2	1.9	2.8	1.0	1.9	-0.2	0.7	-0.6	0.3	0.4	1.3	0.2	1.1		
1ERT (his43)	-0.8			0.2	1.0	0.0	0.8	0.4	1.2								
1DE3 (his137)[c]*	-0.5	2.1	2.6	1.3	1.8	-4.3	-3.8	-1.3	-0.8			0.7	1.2				
3RN3 (his48)*	-0.2	0.9	1.1	4.9	5.1	-3.1	-2.9	0.1	0.3	0.1	0.3			2.5	2.7		
3RN3 (his119)[c]	0.2			-0.5	-0.7	0.2	0.0	1.1	0.9	-0.1	-0.3			-0.9	-1.1		
3RN3 (his12)[c]	-0.3			0.5	0.8	-4.5	-4.2	-0.3	0.0	-0.5	-0.2			-2.1	-1.8		
1DE3 (his104)*	0.2	0.9	0.7	2.3	2.1	0.1	-0.1	-0.4	-0.6			1.3	1.1				
1DE3 (his36)	0.5			0.7	0.2	0.2	-0.3	0.3	-0.2			1.6	1.1				
2RN2 (his62)	0.7			0.7	0.0	0.7	0.0	0.6	-0.1	0.7	0.0	0.7	0.0	0.4	-0.3		
2RN2 (his124)*	0.8	-0.8	-1.6	-1.1	-1.9	0.1	-0.7	0.3	-0.5	-0.5	-1.3			-1.8	-2.6		
1DE3 (his50)*	**1.4**	3.5	2.1	3.1	1.7	-3.8	-5.2	0.4	-1.0			2.1	0.7				
1RGA (his92)[c]	**1.5**			1.2	-0.3	-0.4	-1.9	0.4	-1.1	1.1	-0.4			0.8	-0.7		
1RGA (his40)[c]	**1.6**			1.6	0.0	2.5	0.9	0.1	-1.5	1.1	-0.5			2.7	1.1		
2RN2 (his127)	**1.6**			1.6	0.0	0.8	-0.8	1.1	-0.5	1.3	-0.3			0.7	-0.9		
1DG9 (his66)	**2.0**			1.3	-0.7	1.3	-0.7	1.1	-0.9								
2LZM (his31)*	**2.8**	1.5	-1.3	0.9	-1.9	1.3	-1.5	0.9	-1.9			3.9	1.1				
1DG9 (his72)	**2.9**			3.2	0.3	1.6	-1.3	0.7	-2.2								
RMSD (N)		1.9 (8)		1.8 (20)		2.2 (20)		1.4 (20)		0.5 (9)		1.4 (8)		1.6 (9)		(0)	
Mean Abs Dev		1.8		1.3		1.6		1.1		0.4		1.2		1.4			
Max Abs Dev		2.4		5.1		5.2		3.2		1.3		2.8		2.7			
Total RMSD (N)		1.9 (21)		1.4 (80)		1.4 (80)		1.1 (65)		0.7 (43)		1.2 (32)		1.4 (50)		0.3 (6)	
Total MAD		1.8		1.0		1.0		0.8		0.5		0.9		1.0		0.3	
Total MAX		2.5		5.1		5.2		3.2		1.7		3.3		4.3		0.5	

[a] Root mean squared deviations (rmsd), mean absolute deviation (MAD), maximum absolute deviation (MAX) for predicted pK_a values are shown in red for each residue type. [b] Numbers in parentheses represent the number of values used to calculate rmsd, MAD, MAX. $\Delta pK_a = pK_{a,R} - pK_{a,M}$ from eq 1, where $pK_{a,R}$ is the residue in the protein and $pK_{a,M}$ is the model compound in solution. The values for $pK_{a,M}$ that were used for the model compound for each residue type were the following: Asp 4.0, Glu 4.4, Lys 10.5, His 6.3. Experimental ΔpK_a values in bold indicate residues whose experimental pK_a values vary more than 1 pK_a unit from their "normal" value in solution. PDB codes are color coded as follows: purple, buried residues (accessibility <20%); green, surface residues (accessibility >50%); blue, intermediate residues (20−50%).[17,18,38,39,41,43,45] [c] Catalytic residues. * Included several explicit water molecules as described in the Computational Methods section.

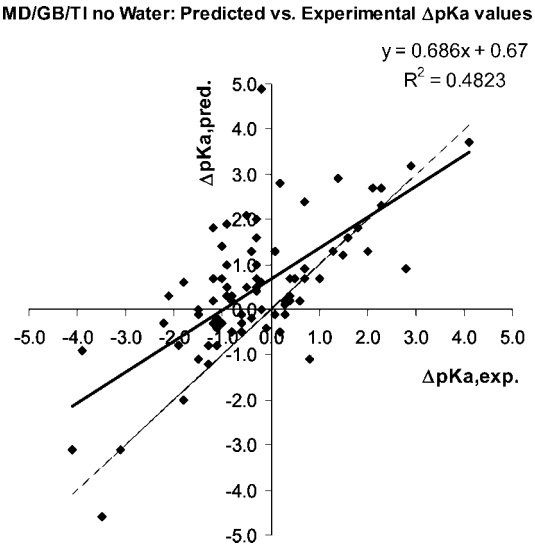

Figure 4. Plot of the predicted ΔpK_a versus experimental ΔpK_a values from Table 1 for the MD/GB/TI method with no explicit waters (under heading "MD/GB/TI w/out water"). ΔpK_a is defined as $\Delta pK_a = pK_{a,R} - pK_{a,M}$ from eq 1. The values used for $pK_{a,M}$ are Asp 4.0, Glu 4.4, Lys 10.5, and His 6.3. The plot contains 80 data points that were fitted linearly with a trendline. The resulting equation is shown with the corresponding R^2 value. The dashed line has a slope of 1 for comparison.

rerun due to the added computational cost. The path dependence of the change in free energy (i.e., forward reaction versus backward reaction) was also explored for two residues: 1A2P Glu60 and 1XNB Asp11. The free energy change was not shown to be significantly path-dependent in either case, with a difference in the free energy changes of <0.2 kcal/mol.

PROPKA. The PROPKA method has a relatively high total rmsd of 1.4 pK_a units compared to the PB methods, but has an equivalent rmsd as compared to the MD/GB/TI method. The breakdown per residue, however, shows that the method works better for Glu and Lys residues with an rmsd of close to 1.0 for both of these residue types. The maximum absolute deviation is also relatively low for Glu and Lys, 2.0 and 2.1 pK_a units, respectively. It fares somewhat worse for Asp with an rmsd of 1.3 and MAX of 4.4 and fares much worse for His with an rmsd of 2.2 and MAX of 5.2.

Figure 5 shows the same type of plot for PROPKA as is shown in Figure 4 for MD/GB/TI. The first plot shows the data for all residue types. The correlation is weaker than for MD/GB/TI, with an R^2 value of 0.28, indicating worse predictive power by PROPKA. The difference between means of the PROPKA data set and the MD/GB/TI data set is statistically significant at the 0.05 level (p-value < 0.0001).

A look at the same data plotted per residue type shows differences in predictions for the different residue types. Glu has the best correlation (R^2 is 0.49), followed by Lys (R^2 is 0.48) and Asp (R^2 is 0.48), and His (R^2 is 0.12), respectively. It appears that, for many of the His cases, the "local desolvation" effects that reflect the degree of burial in an area 4−5 Å surrounding the residue are significantly overestimated and reduce the pK_a considerably more than any of the other contributors that PROPKA estimates (i.e., Coulomb interactions and hydrogen bonding with nearby residues). One of the worst His predictions (His12 of RNase A, PDB code 3RN3), with an error of −4.2 pK_a units, is discussed in the Jensen paper.[18] The pK_a for this residue is known experimentally to be dependent on salt concentration.[56] The lack of explicit interactions between the residue and ions in the solvent is blamed for the poor prediction. The MCCE method also had trouble with this residue, giving an error of −2.0. However, three other methods were able to predict the change in pK_a within 1 unit, so it is unlikely that explicit consideration of salt−residue interactions is required for this residue.

The worst PROPKA prediction for Asp in this benchmark set is Asp75 of barnase (PDB 1A2P), with an error of −4.4 pK_a units. It is the worst Asp prediction in the Jensen study as well. The error in this case is blamed on particularly strong interactions with two nearby Arg residues. It is probably due to the double-counting of interactions with both Arg residues. The EGAD method also had a problem predicting the pK_a of this residue. However, three of the methods were able to predict the change in pK_a to within 1 unit.

A closer look at the plot for Lys predictions reveals that PROPKA tends to predict no change in pK_a for lysine. In fact, 40% of the Lys residues were predicted to have zero change, despite the fact that none of them actually had an experimental change of zero. This tendency can be explained by the observation that PROPKA predicted the majority of lysines to be surface residues. Charge−charge interactions for surface residues are not calculated in the PROPKA method, and few of the lysines in the benchmark set had hydrogen bonding interactions, leading to an unchanged pK_a. This might suggest that the criterion for a lysine residue to be considered buried is too stringent, or perhaps a third category between surface and buried would be useful.

The poor performance in some cases is probably due to lack of sampling, and incorporation of some type of sampling technique would most likely improve PROPKA predictions. The method shows great promise considering it is entirely empirical and the calculations take only seconds. The performance is quite comparable to the MD/GB/TI method, which took between 24 and 48 h per residue on a single Pentium III processor depending on protein size. In this study, PROPKA was used to assign initial protonation states before the MD/GB/TI calculations were performed, since the correct direction of pK_a change was predicted for 82% of the benchmark set (80% for His, which has the highest rmsd).

Other Methods. The geometry-dependent dielectric method of Hellinga and co-workers[39] has a low total rmsd of 1.1 for this benchmark set. It also has the most reported values in common with this benchmark set—65 out of 80. For this reason, a comparison between this method and the PROPKA and MD/GB/TI methods is the most valid. The maximum absolute deviation, 3.2 pK_a units, is comparable to that for MD/GB/TI method (with explicit waters), 2.6, and much lower than for PROPKA, 5.2.

The microenvironment SCP method of Mehler and Guarnieri[38] has the lowest rmsd of the PB methods, 0.7 pK_a units, calculated with 43 of the 80 benchmark residues. It was also the only method that predicted His residues with a better rmsd

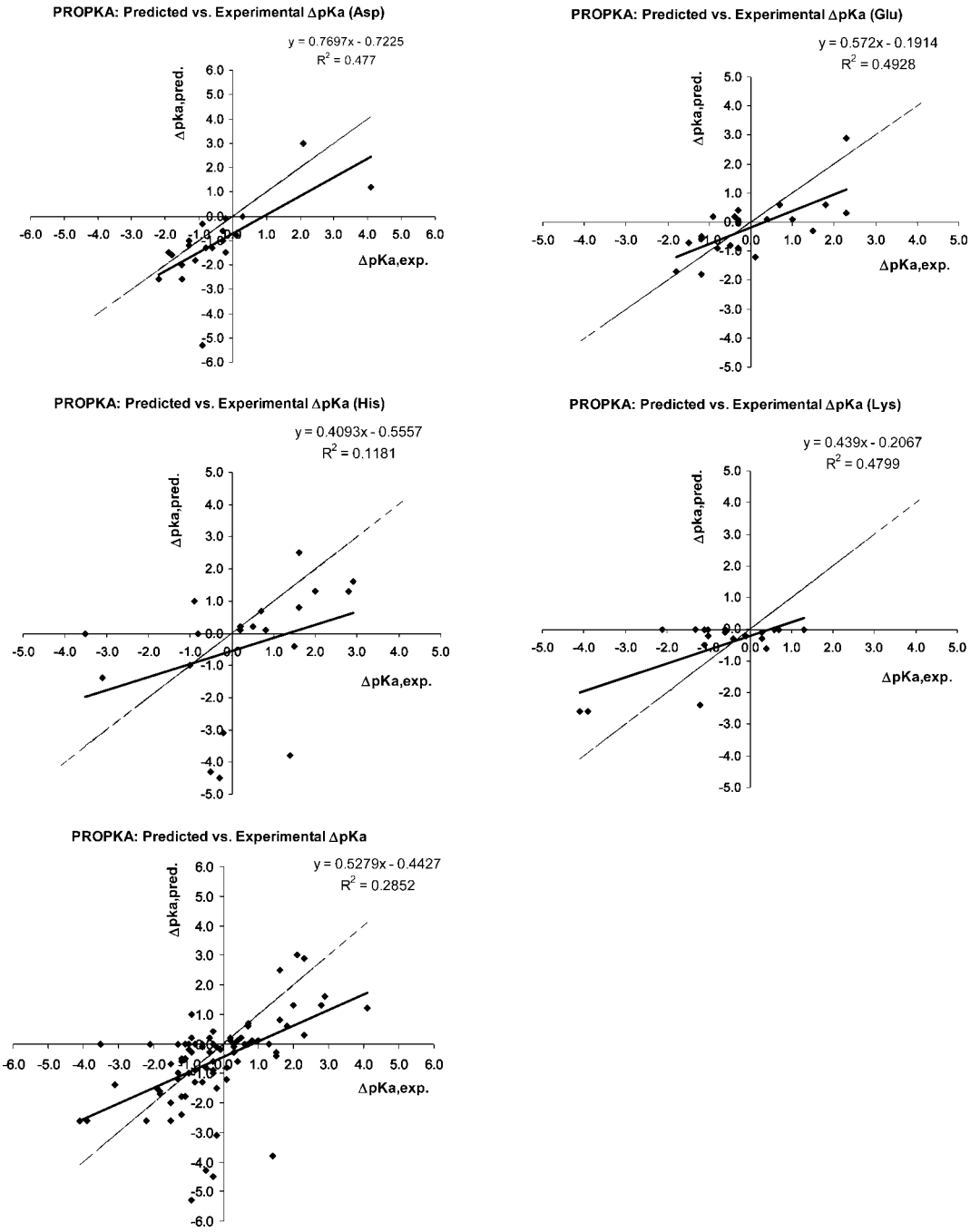

Figure 5. Plots of the predicted ΔpK_a versus experimental ΔpK_a values from Table 1. ΔpK_a is defined as $\Delta pK_a = pK_{a,R} - pK_{a,M}$ from eq 1. The last plot includes all 80 data points, while the remaining plots show the 20 data points for each residue type. In each case, the data were fitted linearly with a trendline. The resulting equation is shown for each graph with the corresponding R^2 value. The dashed line has a slope of 1.

(0.5) and MAX (1.3) than its overall values. The EGAD method and the MCCE method have a comparable total rmsd of 1.2 and 1.4, respectively. These values are slightly higher than their reported values of 0.92 and 0.86 pK_a units. The discrepancy is most likely due to the demanding nature of this benchmark set, which contains a high ratio of buried versus surface residues (21/32). The maximum absolute deviations for EGAD, 3.3, and MCCE, 4.3, are slightly higher than the other PB methods. The QM/MM method has the lowest rmsd, 0.3, and the lowest maximum absolute deviation, 0.5. However, there are only seven entries in Table 1 for this very expensive method, and none of which are buried residues. It would be interesting to see if QM/MM methods can maintain an extremely low rmsd even with buried residues.

C. Variance and SASA. The variance referred to here is the absolute change in pK_a on going from solution to protein. Special care was taken to choose a benchmark set that contains equal numbers of low and high variants, as mentioned in the Introduction. The assumption is that predicting high variants is more difficult, due to the complexity of interactions that cause a large change in pK_a. However, the data from this benchmark test do not show any such relationship. The error was plotted versus the variance (data not shown) giving a very small R^2 value of 0.05.

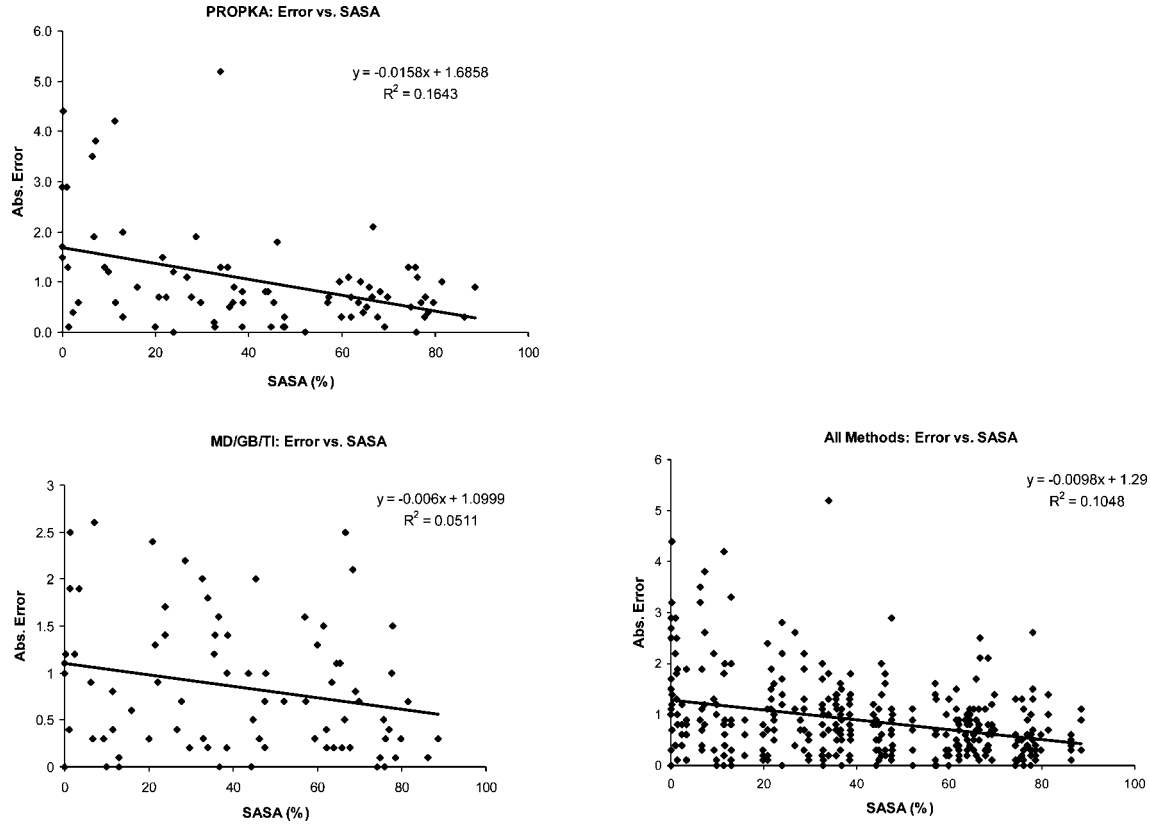

Figure 6. Plots of the calculated error versus the percent of solvent accessible surface area (SASA). (top) Results for the benchmark set including all the methods from Table 1. (bottom, left) Results for the benchmark set from only PROPKA calculations with 80 data points. (bottom, right) Results for the benchmark set from only MD/GB/TI calculations with 80 data points.

Another property that was explored here is the solvent accessible surface area (SASA). The importance of the SASA in determining the transfer energy associated with moving a compound from polar solvent to nonpolar solvent has been well established.[90–92] The thermodynamic cycle in Figure 2 shows why SASA is important for the calculation of pK_a's. As noted in the Introduction, the relative pK_a can be calculated by comparing the change in energy of transferring the protonated and unprotonated residue from solvent to protein. The relationship between error and the percent SASA was explored in Figure 6. It was expected that the error would decrease with increasing SASA, since buried residues have very different interactions than the model compound in solution.

The MD/GB/TI and PROPKA methods were plotted separately, and a third plot contains all the data from Table 1. The plot of all the data shows the expected trend—as the residues are more exposed to solvent (going from left to right on the *x*-axis), the error becomes closer to zero and predicting the change in pK_a becomes easier. The correlation is statistically significant (p-value < 0.0001; i.e., assuming the correlation is just chance, the probability of getting the results we did is less than 0.01%). Despite the statistical significance, the relationship is surprisingly weak, as indicated by a small R^2 value of 0.11. The PROPKA data show a stronger relationship with an R^2 value of 0.16 (p-value < 0.0001), while the MD/GB/TI data shows a significantly weaker relationship with an R^2 value of 0.05 (p-value is 0.02). The fact that SASA is a better predictor of error for PROPKA than for MD/GB/TI makes sense, since the latter method is not parametrized with a set of data that is largely dominated by surface residues, as is the case for PROPKA. When the points corresponding to MD/GB/TI method are removed from the all data plot in Figure 6, the R^2 value is increased to 0.12, showing a stronger relationship between error and SASA. However, the other methods do not collectively show as strong a relationship as PROPKA.

Ionic Strength and pK_a Dependence. One source of error in pK_a prediction is the ability to model accurately the effect of the surrounding ionic strength in solution. While most pK_a's are not highly dependent on ionic strength, one residue in particular that was used in this benchmark is known experimentally to be sensitive to salt concentration. The pK_a of Glu10 in turkey ovomucoid third domain is known to increase by 0.8 pK_a units on going from 1 M KCl to 1 mM KCl.[52] This relationship was probed using the MD/GB/TI method. The pK_a was predicted to increase by 0.7 pK_a units, indicating that this method can accurately model the effects of ionic screening, despite the implicit representation of ions in solution. However, in this case, it is known experimentally that conformational changes and not direct ionic interactions are responsible for the change in pK_a. For the latter interaction, an explicit representation is probably desirable. There is also evidence that the effect of salt concentration on pK_a's cannot be entirely accounted for by an ionic screening model.[93] However, this simple model was sufficient in this example.

D. Case Study: 2-Deoxyribose-5-phosphate Aldolase. Aldolases catalyze stereoselective reactions that involve carbon−carbon bond formation. This characteristic has been

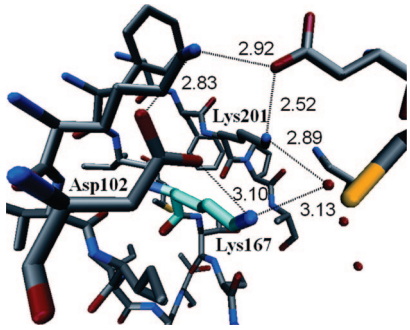

Figure 7. Active site of 2-deoxyribose-5-phosphate aldolase (PDB code: 1P1X). The catalytic lysine (Lys167) is shown in blue. There is an intricate hydrogen bonding network that involves two other lysines, two aspartic acids, and a crystal water.

exploited in the biocatalysis of useful products.[94,95] One well-studied aldolase is 2-deoxyribose-5-phosphate aldolase (DERA). As mentioned in the Introduction (see Figure 1), DERA catalyzes the aldol reaction of acetaldehyde and D-glyceraldehyde 3-phosphate to form 2-deoxyribose-5-phosphate. The first step of the reaction involves nucleophilic attack by an unprotonated lysine. Experimental isolation of the carbinolamine and Schiff base intermediates has proven that Lys167 is the catalytic active site residue.[8] Though the pK_a of Lys167 has not yet been determined experimentally, the pK_a must be close to 7 in order for the reaction to occur. The pK_a was calculated with the MD/GB/TI method to be 6.9 (ΔpK_a = −3.5). PROPKA predicted the pK_a of Lys167 to be 7.7. Figure 7 shows the active site of DERA, with Lys167 highlighted in blue. There is an intricate network of hydrogen bonds involving two aspartic acids, three lysines, and a crystal water. In the crystal structure, Lys167 is hydrogen bonded to one of the aspartates and the water. A neutral lysine would be favored in that position since the nearby aspartates are already ion-paired with other lysines. This case is an example where the MD/GB/TI and PROPKA methods can be used to identify possible catalytic residues.

VI. Conclusions

The benchmark set was chosen to include equal numbers of high (Δp$K_a \geq 1$ from solution to enzyme) and low (Δp$K_a < 1$ from solution to enzyme) variants, with the assumption that it is more difficult to predict high variants since the interactions that produce a change in pK_a are difficult to model. However, the variance was not shown to be a good predictor of the error in pK_a calculations. The SASA proved to be a somewhat better predictor and did show the expected trend of decreasing error with increasing solvent accessibility for all methods except for MD/GB/TI, which did not have a significant correlation.

Most recent methods for predicting pK_a values report an rmsd near 1 pK_a unit. The methods studied here and their corresponding total rmsd values for this benchmark set are the following: the MD/GB/TI method of Simonson et al.[17]—rmsd 1.4 with 80 values; the PROPKA method of Jensen and co-workers[18]—rmsd 1.4 with 80 values; the geometry-dependent dielectric method of Wisz and Hellinga[39]—rmsd 1.1 with 65 values; the microenvironment SCP method of Mehler and co-workers[26]—rmsd 0.7 with 43 values; the EGAD method of Pokala and Handel[41]—rmsd 1.2 with 32 values; the MCCE method of Georgescu et al.[43]—rmsd 1.4 with 50 values; and the QM/MM method of Molina and co-workers[45]—rmsd 0.3 with 6 values. Most of the methods were shown to produce fairly consistent results regardless of residue type, with the exception of His. Most methods fared somewhat worse for this residue type, except for the microenvironment SCP method. This is most likely due to the tautomerism of His, which is not explicitly modeled in any of the methods. A similar increase in rmsd for His residues was found in another pK_a benchmark study.[96]

A much higher MAX than rmsd was found for all methods (except in the case of the QMMM method). The highest MAX value of 5.2 pK_a units was a prediction from the PROPKA method. However, the other more sophisticated PB methods did only marginally better giving MAX values of 3.2,[39] 3.3,[41] and 4.3.[35] The fully atomistic MD/GB/TI method was shown to be equally unimpressive in this measure, giving MAX values of 2.5 and 5.1, with and without explicit waters, respectively (with longer simulation time, the MAX was reduced to 3.9 without explicit waters). The only method that gave a MAX less than 2 was the microenvironment SCP, which seems to be the most promising method. However, considering the high absolute errors, the challenge remains to find a highly accurate method that can be done with minimal computational cost. Improvements may include constant longer simulation times, pH simulations in which the pK_a of ionizable residues can change,[97] incorporation of cofactors and metal ions, improved solvent models and inclusion of some explicit water, and improved force fields.

Acknowledgment. We are grateful for the comments, criticisms, and helpful discussions contributed by David Case, Ernest Mehler, Navin Pokala, Tracy Handel, and especially Jan Jensen during the evolution and progress of this work. We are grateful to the Defense Advanced Research Projects Agency (DARPA) for financial support of this research. Part of this work was performed with funding from a University of California Lawrence Livermore National Laboratory (LLNL) graduate fellowship to C.S. under Contract No. B558556.

References

(1) Hendsch, Z. S.; Jonsson, T.; Sauer, R. T.; Tidor, B. *Biochemistry* **1996**, *35*, 7621–7625.

(2) Elcock, A. H.; McCammon, J. A. *J. Mol. Biol.* **1998**, *280*, 731–748.

(3) Schaefer, M.; Sommer, M.; Karplus, M. *J. Phys. Chem. B* **1997**, *101*, 1663–1683.

(4) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159.

(5) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284–290.

(6) Warshel, A. *Biochemistry* **1981**, *20*, 3167–3177.

(7) Marshall, S. A.; Morgan, C. S.; Mayo, S. L. *J. Mol. Biol.* **2002**, *316*, 189–199.

(8) Heine, A.; Luz, J. G.; Wong, C. H.; Wilson, I. A. *J. Mol. Bio.* **2004**, *343*, 1019–1034.

(9) Heine, A.; DeSantis, G.; Luz, J. G.; Mitchell, M.; Wong, C.-H.; Wilson, I. A. *Science* **2001**, *294*, 369–374.

(10) Antosiewicz, J.; McCammon, J.; Gilson, M. *Biochemistry* **1996**, *35*, 7819–7833.

(11) York, D. M.; Darden, T. A.; Pedersen, L. G. *J. Chem. Phys.* **1993**, *99*, 8345–8348.

(12) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509–521.

(13) Gilson, M. K.; Rashin, A.; Fine, R.; Honig, B. *J. Mol. Biol.* **1985**, *184*, 503–516.

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

(15) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.

(16) Raquet, X.; Lounnas, V.; Lamotte-Brasseur, J.; Frere, J.; Wade, R. *Biophys. J.* **1997**, *73*, 2416–2426.

(17) Simonson, T.; Carlsson, J.; Case, A. D. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(18) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704–721.

(19) Warshel, A.; Sussman, F.; King, G. *Biochemistry* **1986**, *25*, 8368–8372.

(20) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415–436.

(21) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373–17387.

(22) Warshel, A. *Nature* **1987**, *330*, 15–16.

(23) Warshel, A.; Russell, S. T.; Churg, A. K. *Proc. Natl. Acad. Sci. USA* **1984**, *81*, 4785–4789.

(24) Sham, Y. Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744–1753.

(25) Garcia-Moreno, E. B. *Methods Enzymol.* **1994**, *240*, 645–667.

(26) Hassan, S. A.; Guarnieri, F.; Mehler, E. L. *J. Phys. Chem. B* **2000**, *104*, 6490–6498.

(27) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721–1733.

(28) Gilson, M. K.; Honig, B. H. *Nature* **1987**, *330*, 84–86.

(29) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Chem.* **1990**, *19*, 301–332.

(30) Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins* **2004**, *15*, 252–265.

(31) Pethig, R. *Annu. Rev. Phys. Chem.* **1992**, *43*, 177–205.

(32) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366–4377.

(33) Sternberg, M. J.; Hayes, F. R.; Russell, A. J.; Thomas, P. G.; Fersht, A. R. *Nature* **1987**, *330*, 86–88.

(34) Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11*, 243–252.

(35) Gunner, M. R.; Alexov, E. *Biochim. Biophys. Acta* **2000**, *1458*, 63–87.

(36) Karp, D. A.; Gittis, A. G.; Stahley, M. R.; Fitch, C. A.; Stites, W. E.; Garcia-Moreno, B. E *Biophys. J.* **2007**, *92*, 2041–2053.

(37) Schutz, C.; Warshel, A. *Proteins* **2001**, *44*, 400–417.

(38) Mehler, E.; Guarnieri, F. *Biophys. J.* **1999**, *77*, 3–22.

(39) Wisz, M. S.; Hellinga, H. W. *Proteins* **2003**, *51*, 360–377.

(40) Archontis, G.; Simonson, T. *Biophys. J.* **2005**, *88*, 3888 3904.

(41) Pokala, N.; Handel, T. M. *Protein Sci.* **2004**, *13*, 925–936.

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

(43) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731–1748.

(44) Merz, K. M. *J. Am. Chem. Soc.* **1991**, *113*, 3572–3575.

(45) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

(46) Riccardi, D.; Schaefer, P.; Yang, Y.; Haibo, Y.; Ghosh, N.; Prat-Resina, X.; Koenig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.

(47) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926–935.

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

(49) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180–1184.

(50) Kuramitsu, S.; Hamaguchi, K. *J. Biochem.* **1980**, *87*, 1215–1219.

(51) Oda, Y.; Yoshida, M.; Kanaya, S. *J. Biol. Chem.* **1993**, *268*, 88–92.

(52) Oda, Y.; Yamazaki, T.; Nagayama, K.; Kanaya, S.; Kuroda, Y.; Nakamura, H. *Biochemistry* **1994**, *33*, 5275–5284.

(53) Scaller, W.; Robertson, A. D. *Biochemistry* **1995**, *34*, 4714–4723.

(54) Swint-Kruse, L.; Robertson, A. D. *Biochemistry* **1995**, *34*, 4724–4732.

(55) Gooley, P. R.; Keniry, M. A.; Dimitrov, R. A.; Marsh, D. E.; Gayler, K. R.; Grant, B. R. *J. Biol. NMR* **1998**, *12*, 523–534.

(56) Khare, D.; Alexander, P.; Antosiewicz, J.; Bryan, P.; Gilson, M.; Orban, J. *Biochemistry* **1997**, *36*, 3580–3589.

(57) Baker, W.; Kintanar, A. *Arch. Biochem. Biophys.* **1996**, *327*, 189–199.

(58) Perez-Canadillas, J. M.; Campos-Olivas, R.; Lacadena, J.; del Pozo, A. M.; Gavilanes, J. G.; Santoro, J.; Rico, M.; Bruix, M. *Biochemistry* **1998**, *37*, 15865–15876.

(59) Qin, J.; Clore, G. M.; Gronenborn, A. M. *Biochemistry* **1996**, *35*, 7–13.

(60) Oliveberg, M.; Arcus, V. L.; Fersht, A. R. *Biochemistry* **1995**, *34*, 9424–9433.

(61) Norton, R. S.; Cross, K.; Braach-Maksvytis, V.; Wachter, E. *Biochem. J.* **1993**, *93*, 45–551.

(62) Giletto, A.; Pace, C. N. *Biochemistry* **1999**, *38*, 13379–13384.

(63) Inagaki, F.; Kawano, Y.; Shimada, I.; Takahashi, K.; Miyazawa, T. *J. Biochem.* **1981**, *89*, 1185–1195.

(64) Chen, H. A.; Pfuhl, M.; McAlister, M. S. B.; Driscoll, P. C. *Biochemistry* **2000**, *39*, 6814–6824.

(65) Joshi, M. D.; Hedberg, A.; McIntosh, L. P. *Protein Sci.* **1997**, *6*, 2667–2670.

(66) Garcia-Moreno, E. B.; Dwyer, J. J.; Gittis, A. G.; Lattman, E. E.; Spencer, D. S.; Stites, W. E. *Biophys. Chem.* **1997**, *64*, 211–224.

(67) Damblon, C.; Raquet, X.; Lian, L.-Y.; Lamotte-Brasseur, J.; Fonze, E.; Charlier, P.; Roberts, G. C. K.; Frere, J. M. *Proc. Natl. Acad. Sci.* **1996**, *93*, 1747–1752.

(68) Harris, T. K.; Wu, G.; Massiah, M. A.; Mildvan, A. S. *Biochemistry* **2000**, *39*, 1655–1674.

(69) Lund-Katz, S.; Mohamed, Z.; Wehrli, S.; Dhanasekaran, P.; Baldwin, F.; Weisgraber, K. H.; Phillips, M. C. *J. Biol. Chem.* **2000**, *275*, 34459–34464.

(70) Zhang, G.; Mazurkie, A. S.; Dunaway-Mariano, D.; Allen, K. N. *Biochemistry* **2002**, *41*, 13370–13377.

(71) Lund-Katz, S.; Wehrli, S.; Zaiou, M.; Newhouse, Y.; Weisegraber, K. H.; Phillips, M. C. *J. Lipid Res.* **2001**, *42*, 984–901.

(72) Guanghua, G.; DeRose, E. F.; Kirby, T. W.; London, R. E. *Biochemistry* **2006**, *45*, 1785–1794.

(73) Kesvatera, T.; Jonsson, B.; Thulin, E.; Linse, S. *J. Mol. Biol.* **1996**, *259*, 828–839.

(74) Inagaki, F.; Miyazawa, T.; Hori, H.; Tamiya, N. *Eur. J. Biochem.* **1978**, *89*, 433–442.

(75) Fujii, S.; Akasaka, K.; Hatano, H. *J. Biochem.* **1980**, *88*, 798–796.

(76) Lee, K. K.; Fitch, C. A.; Garcia-Moreno, B. *Protein Sci.* **2002**, *11*, 1004–1016.

(77) Foreman-Kay, J. D.; Clore, G. M.; Gronenborn, A. M. *Biochemistry* **1992**, *31*, 3442–3452.

(78) Dillet, V.; Van Etten, R. L.; Bashford, D. *J. Phys. Chem. B* **2000**, *104*, 11321–11333.

(79) Dao-pin, S.; Anderson, D. E.; Baase, W. A.; Dahlquist, F. W. I.; Matthews, B. W. *Biochemistry* **1991**, *30*, 11521–11529.

(80) Anderson, D. E.; Becktel, W. J.; Dahlquist, F. W. *Biochemistry* **1990**, *29*, 2403–2408.

(81) PROPKA can be accessed via the web at http://propka.ki.ku.dk.

(82) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, 2004.

(83) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383–394.

(84) Hummer, G.; Szabo, A. *J. Chem. Phys.* **1996**, *105*, 2004–2010.

(85) Fraczkiewicz, R.; Braun, W. *J. Comput. Chem.* **1998**, *19*, 319–333.

(86) GETAREA can be accessed via the web at http://pauli.utmb.edu/cgi-bin/get_a_form.tcl.

(87) The GETAREA manual can be accessed via the web at http://pauli.utmb.edu/getarea/area_man.html.

(88) Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; Garcia-Moreno, B. E. *Biophys. J.* **2002**, *82*, 3289–3304.

(89) All linear regression and correlation was done with Microsoft Excel.

(90) Street, A. G.; Mayo, S. L. *Fold Des.* **1998**, *3*, 253–258.

(91) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(92) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(93) Kao, Y.-H.; Fitch, C. A.; Bhattacharya, S.; Sarkisian, C. J.; Lecomte, J. T. J.; Garcia-Moreno, B. E. *Biophys. J.* **2000**, *79*, 1637–1654.

(94) Gijsen, H. J. M.; Qiao, L.; Fitz, W.; Wong, C. H. *Chem. Rev.* **1996**, *96*, 443–473.

(95) Seebeck, F. P.; Guainazzi, A.; Amoreira, C.; Baldridge, K. K.; Hilvert, D. *Angew. Chem., Int. Ed.* **2006**, *45*, 6824–6826.

(96) Davies, M. N.; Toseland, C. P.; Moss, D. S.; Flower, D. R. *BMC Biochem.* **2006**, *7*, 18–30.

(97) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

CT8000014