FEATURE ARTICLE

Prediction and Rationalization of Protein pK_a Values Using QM and QM/MM Methods

Jan H. Jensen,*,†,‡ Hui Li,†,‡,§ Andrew D. Robertson,⊥ and Pablo A. Molina†,||

Department of Chemistry, Center for Biocatalysis and Bioprocessing, and Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242

Received: April 13, 2005; In Final Form: May 12, 2005

We describe the development and application of a computational method for the prediction and rationalization of pK_a values of ionizable residues in proteins, based on ab initio quantum mechanics (QM) and the effective fragment potential (EFPs) method (a hybrid QM/MM method). The theoretical developments include (1) a covalent boundary method based on frozen localized orbitals, (2) divide-and-conquer methods for the ab initio computation of protein EFPs consisting of multipoles up to octupoles and dipole polarizability tensors, (3) a method for computing vibrational free energies for a localized molecular region, and (4) solutions of the polarized continuum model of bulk solvation equations for protein-sized systems. The QM-based pK_a prediction method is one of the most accurate methods currently available and can be used in cases where other pK_a prediction methods fail. Preliminary analysis of the computed results indicate that many pK_a values (1) are primarily determined by hydrogen bonds rather than long-range charge−charge interactions and (2) are relatively insensitive to large-scale dynamical fluctuations of the protein structure.

Introduction

From a molecular point of view the uptake or release of a proton from the solvent is one of the simplest chemical reactions in aqueous solution. However, it is also one of the most important reactions because it determines the pH dependence of the molecular charge, a key determinant of the chemical and physical properties of the molecule. It is therefore crucial to know the pK_a values of ionizable functional groups (i.e., groups with pK_a values in the pH range of interest) in molecules.

Proteins, for example, almost always contain amino acids with ionizable groups, which are important for intraprotein, protein−solvent, and protein−ligand interactions,[1] and play key roles in protein solubility, folding, stability, binding ability, and catalytic activity. The pK_a values of the ionizable residues are thus the basis for understanding the pH-dependent characteristics of proteins and catalytic mechanisms of many enzymes.[2]

Although there are only a few different kinds of ionizable functional groups (−COOH, −SH, phenol, $-NH_3^+$, imidazolium, and guanidinium) in proteins, their pK_a values can be significantly affected by their location in the protein structure. For example, two different −COOH groups in the same protein can have respective pK_a values of 2 and 9 pH units, implying significantly different chemical environments.[3]

pK_a values must therefore be obtained for each individual ionizable residue, typically by measuring some spectroscopic property of the residue as a function of pH. With the exception of cysteine and tyrosine residues (−SH and phenolic-OH, which are UV/vis active), the only practical method is NMR. However, the assignment of NMR chemical shifts is nontrivial for small proteins (<∼200 residues) and extremely difficult for proteins of average size (∼300−500 residues). Despite their importance, pK_a values are thus not known for most proteins of interest.

The theoretical prediction of protein pK_a values based on protein structures has therefore been a central challenge to biomolecular modeling for many years. Virtually all pK_a prediction methods for proteins[4−9] are based on

$$pK_a = pK_{model} + [\Delta G_{env}(A^-) - \Delta G_{env}(HA)]/1.36 \quad (1)$$

[1.36 = RT ln(10) at 298 K in kcal/mol] whereby the protein pK_a value is related to the experimentally determined pK_a value of a model compound (pK_{model}), shifted by the change in the "environment energy" ($\Delta\Delta G_{env}$). ΔG_{env} contains contributions from desolvation and interactions of the ionizable group with the rest of the protein. The methods differ mainly in how ΔG_{env} is calculated and whether the protein geometry remains fixed during the calculation.

Tanford and Kirkwood[10] developed one of the earliest methods, in which only interactions between charged groups (represented as point charges embedded in spherical continuum medium with a low dielectric) are considered. The spherical shape allows for an analytic solution to the electrostatic (Poisson−Boltzmann) equation.

One problem with the Tanford−Kirkwood approach is that other interactions, such as hydrogen bonding, are known to affect protein pK_a values. Thus, in general, a fully atomistic description of the protein is needed and for most methods the interaction energy is calculated as the interaction of the atomic charges (from a MM force field) in the model system with the

* Corresponding author. E-mail: jan-jensen@uiowa.edu.
† Department of Chemistry.
‡ Center for Biocatalysis and Bioprocessing.
⊥ Department of Biochemistry.
§ Current address: Department of Chemistry, Iowa State University, Ames, IA 50011.
|| Current address: Department of Chemistry, Murray State University, Murray, KY 42071.

10.1021/jp051922x CCC: $30.25
Published on Web 07/12/2005

Jan H. Jensen was born in Denmark in 1969 and came to the U.S. as a foreign exchange student in 1985. He received his B.A. in Chemistry from Concordia College in 1989 and his Ph.D. in Theoretical Chemistry from Iowa State University in 1995, working with Mark Gordon. He continued in the Gordon group as a postdoctoral associate until 1997, when he moved to the University of Iowa where he is currently Associate Professor of Chemistry. His research interests are primarily in the area of computational molecular biophysics—at the intersection of molecular physics, quantum chemistry, and structural biology/bioinformatics.

Hui Li received his B.S. in chemistry from Lanzhou University in 1993 and his Ph.D. in quantum/theoretical chemistry from the University of Iowa in 2004. He is currently a postdoctoral research associate at Iowa State University. His research interests include the developments of quantum/molecular mechanical methods and their applications in modeling molecular interactions, solvent effects, and biomolecular properties.

Andrew Robertson received a B.A. in Biology from the University of California San Diego in 1981. After a year as a laboratory technician with John S. OBrien in the UC San Diego School of Medicine, he pursed the Ph.D. in Biochemistry at the University of Wisconsin Madison with John L. Markley. Three years of postdoctoral training with Robert L. Baldwin in Biochemistry at Stanford University was followed by a faculty position in Biochemistry at the University of Iowa College of Medicine, where he has been since 1991. His research interests focus in the area of structure-energetics relationships in proteins.

Pablo A. Molina received his Ph.D. in Chemistry from the University of Iowa with Jan Jensen in 2003 along with a Masters in Science Education and a Masters in Latin. He has been an Assistant Professor at Murray State University for the past two years. His primary research interests are hydrogen bonding interactions in aspartyl proteases and the study of pK_a values of ionizable residues.

electrostatic potential of the protein. The most popular prediction methods obtain the electrostatic potential of the protein and solvent by numerically solving the linearized Poisson−Boltzmann equation (LPBE).[4−6,8,11−15] Approaches based on protein dipole−Langevin dipole,[16] linear response approximation,[16,17] free energy perturbation methods,[18,19] and screened Coulomb potentials[9,20,21] have also been used. In the MM/LPBE approach the dielectric constant of the bulk solvent region is 80, and a lower (typically 4 or 20) dielectric constant is used for the protein interior.[12,22,23]

We recently compared published results for five state-of-the-art pK_a prediction methods applied to the prediction of 83 pK_a values in five different proteins.[24] The root-mean-square-deviation (RMSD) from experiment ranged from 0.6 to 1.1 pH units, and the largest absolute error for each method ranged from 1.7 to 4.3 pH units. As noted before,[23] the methods tend to perform best for residues on the protein surface, which tend to have pK_a values close to pK_{model}. Because most ionizable residues are on the protein surface, this results in relatively low RMSDs from experiment. However, functionally important residues tend to reside in the protein interior and exhibit more extreme pK_a values that are harder to predict accurately.[23]

One possible limitation of current methods is the accuracy of the treatment of short-range interactions such as hydrogen bonds by the molecular mechanics force fields. There is mounting evidence that an atom-centered charge (ACC) model is not always an adequate representation of the molecular electrostatic potential.[25−37] For example, the ACC model tends to underestimate the directionality of hydrogen bonds,[34] whereas models that include additional charges[34] or higher-order multipoles[25,26] reproduce ab initio results significantly better.

Several "multipole libraries" have been or are being developed for amino acids,[30,38−41] and at least two force fields[33,42] employ a multipole-based electrostatic model. However, the transferability of the multipole parameters can be complicated by the conformational dependence of the higher moments and, more generally, because a higher degree of accuracy (compared to charge-based models) is typically sought.[30,37,41,43,44] These next-generation force fields have not yet been applied to the prediction of protein pK_a values.

Alternatively, better pK_a predictions may be obtained by treating the short-range interactions involving ionizable residues with ab initio quantum mechanics (QM) and the long-range interactions with classical (molecular mechanics, MM) electrostatic techniques using a hybrid QM/MM approach. Here we describe our work on this approach during the last seven years.

The organization of the paper largely reflects the order in which we approached the problem: We first established a QM-based pK_a prediction method that gave good results for small molecules. Then we developed several algorithms that allowed us to calculate the corresponding energy components (deprotonation and solvation energies) at roughly the same level of theory for protein-sized systems using QM/MM. With these algorithms in hand we demonstrated, for the first time, that a protein pK_a value could be accurately predicted by a QM/MM method. Following this proof-of-concept study, we improved the efficiency of some of the algorithms, most notably the calculation of solvation energies, and tested the sensitivity of the pK_a values to various aspects of our model. This work led us to develop a significantly simpler and more efficient pK_a prediction method. The simplicity of the method greatly facilitated the determination of the key pK_a determinants, which subsequently let us propose a set of simple rules by which the effect of protein structure on pK_a values can be understood and

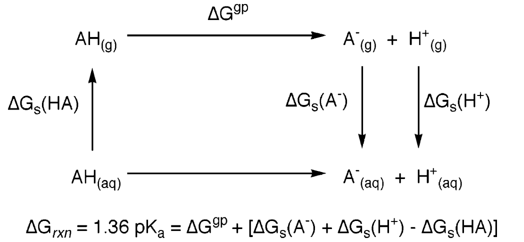


Figure 1. Thermodynamic cycle relating the pK_a to the gas-phase proton basicity ΔG^{gp}) via the solvation energies [ΔG_s] of the products and reactants. The value 1.36 corresponds to RT ln(10) at 298 K in kcal/mol.

quantitatively predicted. We conclude by discussing the larger implications of our work as well as future directions.

Small Molecule pK_a Predictions

Methodology. On the basis of previous work,[45−58] we developed the following methodology for small molecule pK_a predictions (cf. Figure 1).[59] The gas-phase basicity, ΔG^{gp}, is calculated at the MP2/6-31+G(2d,p)//RHF/6-31G(d) level of theory,[60] using frequencies scaled by 0.89 for the vibrational free energy correction (ΔG^{vib}; ΔG^{trans} and ΔG^{rot} are the translation and rotational free energies, respectively),

$$\Delta G^{gp} = \Delta E^{MP2//RHF} + \Delta G^{trans} + \Delta G^{rot} + \Delta G^{vib} + RT\ln(\tilde{R}T) \quad (2)$$

The last term in eq 2 changes the reference state from 1 atm to 1 M [$K_c = K_p(\tilde{R}T)$ for AH → $A^- + H^+$ reactions, where $\tilde{R}$ = 0.082 06 (L·atm)/(mol·K)].[57]

The solvation energy is calculated by the default solvation method in the Gaussian98 program, which is Tomasi's polarized continuum model (PCM) using the united atom for Hartree−Fock (UAHF) radii proposed by Barone, Cossi, and Tomasi,[61] and gas-phase geometries. The UAHF model is a set of rules, based on atomic number, connectivity, and charge, for determining radii of the spheres used to define the solute/solvent boundary. The rules are determined empirically so that they reproduce experimental solvation energies for small molecules. The electrostatic contribution to the solvation energy is calculated using the dielectric PCM (D-PCM)[62] and the ICOMP = 4 charge normalization procedure,[63] whereas the dispersion−repulsion and cavitation contributions are calculated by the methods of Floris et al.[64] and Pierotti,[65] respectively (these are default options in the Gaussian98 program),

$$\Delta G_s(X) = \Delta G_{elec}(X) + \Delta G_{cav}(X) + \Delta G_{disp-rep}(X) \quad (3)$$

The UAHF parametrization was done at the RHF/6-31G(d) level of theory for neutral molecules and cations and RHF/6-31+G(d) for anions, using gas-phase geometries. This study uses the same level of theory for the calculation of the solvation energy, except that RHF/6-31+G(d)//RHF/6-31G(d) is used for anions to avoid optimizing geometries twice.

Following Topol et al.,[48] we use a value of −262.5 kcal/mol for $\Delta G_s(H^+)$.

Results and Discussion. The predicted pK_a values for acetic acid, methylamine, imidazole, phenol, and ethanethiol (small-molecule models of common ionizable groups in proteins) are listed in Table 1, together with the energy components defined in eqs 2 and 3. The D-PCM/ICOMP = 4 procedure results in pK_a values that are within 0.9 pH units of experiment with a root-mean-square deviation (RMSD) of 0.6 pH units. All predicted pK_a values are underestimated, and the RMSD can be decreased to 0.3 pH units by increasing $\Delta G_s(H^+)$ to −261.8 kcal/mol, which is well within the range of experimental estimates.[9] Comparisons of predictions of previously published methods indicate that the method proposed here is at least as accurate, and often better.

The ICOMP = 4 charge renormalization method is not available in the GAMESS program[66] (which we will use for the protein pK_a predictions), so the use of the integral equation formalism[67] PCM (IEF-PCM) without charge renormalization (ICOMP = 0) is investigated here. The use of IEF-PCM/ICOMP = 0 for the calculation of the solvation energies leads to essentially unchanged results for methylamine and imidazole (Table 1). Larger errors are observed for acetic acid, phenol, and methanethiol, because charge penetration into the continuum is more pronounced for anions. Thus, for the calculation of protein pK_a values it will be necessary to estimate the effect of D-PCM/ICOMP = 4 on the ionizable residue in question. This is accomplished by the ONIOM-PCM/X method, as described below.

Effective Fragment Potential Method

The effective fragment potential (EFP) method,[68,69] is a hybrid QM/MM method in which only the active part of a molecular system is treated with ab initio quantum mechanics and the rest is replaced by one or more EFPs. An EFP represents the static electrostatic potential by a distributed multipole expansion[27] (charges through octupoles at all atomic centers and bond midpoints), whereas the electronic polarizability is represented by dipole polarizability tensors for each valence (localized) molecular orbital.[36] The main feature that distinguishes the EFP approach from other QM/MM methods is that the EFP is generated from other QM calculations and does not contain any adjustable parameters. Thus, in principle, the quality of the EFP and, hence, the QM/EFP Hamiltonian is systematically improvable, in analogy with conventional electronic structure theory.

TABLE 1: Computed and Experimental pK_a's of Small Molecules with Functional Groups Found in Amino Acid Residues[a]

			D-PCM/ICOMP = 4		IEF-PCM/ICOMP = 0		
acid	ΔE^{MP2} [b]	ΔG_{trv} [c]	$\Delta\Delta G_s$ [d,f]	pK_a	$\Delta\Delta G_s$[d]	pK_a	exp
acetic acid	352.55	−13.28	−333.08	4.6	−330.81	6.2	4.8
methylamine	223.28	−13.25	−196.72	9.8	−196.76	9.8	10.6
imidazole	231.26	−12.66	−210.34	6.1	−210.19	6.2	7.0
phenol	354.43	−11.99	−329.20	9.7	−322.88	14.4	10.0
methanethiol	360.71	−10.02	−336.90	10.1	−331.85	13.8	10.3
RMSD[e]				0.6		2.6	
Lys55	249.38	−12.26	−221.78	11.3	−223.61	9.9	11.1

[a] The individual energy components used to compute the pK_a's [Figure 1 and eq 2] are also given, in kcal/mol. [b] Gas-phase (electronic) deprotonation energy: $\Delta E^{MP2//RHF}$, cf. eq 2. [c] Gas-phase free energy correction using 1 M reference state: sum of the last four terms in eq 1. [d] Change in solvation energy: last three terms in the equation in Figure 1. [e] Root-mean-square deviation from experiment. [f] Calculated using D-PCM-X/ICOMP = 4, cf. eq 3.

The EFP method was originally developed for the study of discrete solvation effects,[70−74] and several methodological advances were thus necessary to apply the method to protein pK_a predictions.

Covalent Boundary. A method to treat the covalent boundary between QM and EFP regions has been developed by Kairys and Jensen[75] in which a "buffer region" consists of localized molecular orbitals (LMOs) that are kept frozen during the SCF.

Protein EFPs. A divide-and-conquer method for constructing EFPs for proteins has been developed by Minikis, Kairys, and Jensen.[36] In this approach the protein is divided into smaller overlapping pieces, for which a multipole expansion can be generated ab initio, and then reassembled by excluding parameters from the region of overlap.

Free Energies. A vibrational analysis for partially optimized systems that yields accurate free energy changes has been developed by Li and Jensen.[76] In this method only a subset of the atoms (in our case the atoms in the ab initio region) are displaced during a numerical Hessian calculation, resulting in a "partial Hessian". Our study shows that vibrational energy and entropy changes for proton abstraction reactions calculated using frequencies obtained in this manner are within 0.2 kcal/mol of conventional values.

Solvation. The EFP interface with Tomasi's polarized continuum method for treating bulk solvation, developed by Bandyopadhyay et al.,[77,78] was extended to protein-sized systems by Li et al.[59] by decreasing the memory requirement and parallelizing the code.

Prediction of pK_a Values in a Polyprotic Acid

Almost all proteins contain several (n) ionizable amino acids. If these ionizable residues have a nonnegligible interaction energy, then their pK_a values will be interdependent. In principle, the pK_a of a particular site (i) is obtained by determining the pH value at which the protonation probability, θ_i, is 0.5.[7] However, the evaluation of θ_i requires the free energies for all sites for all possible (2^n) protonation states,

$$\theta_i = \langle x_i \rangle = \frac{\sum_j^{2^n} x_i^j \mathrm{e}^{-G_j/RT}}{\sum_j^{2^n} e^{-G_j/RT}} \tag{4}$$

Here x_i^j is 1 or 0 depending on whether site i is protonated or unprotonated in protonation state j, respectively, and G_j is the free energy of protonation state j at a given pH. Several techniques, such as Monte Carlo sampling, have been developed for Poisson−Boltzmann-based methods to reduce the number of energy evaluations, but the general approach is still too time-consuming when the energies are evaluated using QM/MM.

However, the main intent of the QM/MM pK_a prediction method is to rationalize unusual pK_a values that have already been identified experimentally and for which MM/LPBE pK_a predictions already have been (or easily could be) performed. Thus, the calculated or measured pK_a values and site−site interactions can be used to determine the optimum protonation state at a given pH.

Here we use Lys55 in the protein turkey ovomucoid third domain as an example. All pK_a values have been determined experimentally and through classical Poisson−Boltzmann calculations, which also provides the predicted interaction between sites (W).[79]

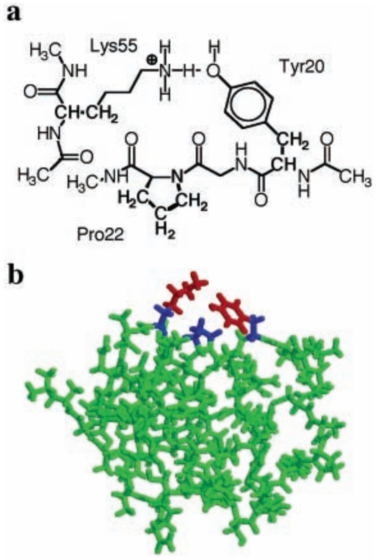


Figure 2. (a) Subsystem of OMTKY3 used to obtain the buffer region (bold) used for (b) ab initio/buffer/EFP regions (red/blue/green) used for the computation of the pK_a of Lys55.

The MM/LPBE calculations predict that the pK_a of Lys55 is affected appreciably (i.e., $|W| > 0.1$ pH units) by only three residues, Tyr20, His52, and CysC56, so that any protonation state can be used for the remaining residues. The respective experimental pK_a values of His52 and CysC56 residues are >8.6 and 3.6 pH units lower than that of Lys55. Both are therefore >99.999% deprotonated at pH = 11.1 and can be treated as 100% deprotonated for the calculation of the pK_a of Lys55. A similar conclusion is reached by using the apparent pK_a values from the MM/LPBE calculations.

Finally, the experimental pK_a values of Tyr20 and Lys55 are equal and are predicted to interact by $W = 0.6$ pH units. Thus, in computing the pK_a value of Lys55, we should consider both protonated and deprotonated Tyr20; i.e., we need to compute energies of four protonation states, which is manageable. (In practice deprotonaing Tyr20, which is included in the ab initio region, results in spontaneous proton transfer from Lys55.)

After our proof-of-concept study, Nielsen and McCammon used a similar approach to compute pK_a values of select residues using Poisson−Boltzmann calculations.[80]

Proof of Concept: Prediction of the pK_a Value of Lys55 Using QM/MM

Methodology. The solution structure of OMTKY3 has been determined using NMR by Hoogstraten et al.[81] and was obtained from the Protein Data Bank (entry 1OMU). We use the first of the 50 conformers without further refinement of the overall structure.

(1) The electronic and geometric structures of the Lys55 and Tyr20 side chains are treated quantum mechanically at the MP2/6-31+G(2d,p)//RHF/6-31G(d) level of theory (Figure 2), and the rest of the protein is treated with an EFP, described in more detail below. The use of the diffuse functions on atoms near the buffer region causes SCF convergence problems due to couplings with the induced dipoles in the EFP region, so the 6-31+G(2d,p) basis set was used only for the $C^{\delta}H_2C^{\epsilon}H_2$-$NH_3$···HO−$C^{\xi}(C^{\epsilon 1,2}H)_2$ atoms in the MP2 calculation.

(2) The ab initio region is separated from the protein EFP by a buffer region[75] composed of frozen localized molecular orbitals (LMOs) corresponding to all the bond LMOs connecting the **bold** atoms in Figure 2a, as well as the core and lone pair LMOs

belonging to those atoms. The Pro22 buffer is needed to describe its short-range interactions with Tyr20.[33] The buffer LMOs are generated by an RHF/6-31G(d) calculation on a subset of the system (shown in Figure 2a), projected onto the buffer atom basis functions, and subsequently frozen in the EFP calculations by setting select off-diagonal MO Fock matrix elements to zero. The ab initio/buffer region interactions are calculated ab initio and thus include short-range interactions.

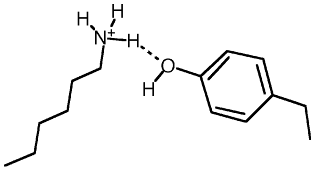
(3) The EFP describing the rest of the protein is generated by nine separate ab initio calculations on overlapping pieces of the protein truncated by methyl groups. Two different regions of overlap are used depending on whether it occurs on the protein backbone or on a disulfide bridge, as described in ref 36. The electrostatic potential of each protein piece is expanded in terms of multipoles through octupoles centered at all atomic and bond midpoint centers using Stone's distributed multipole analysis.[27] The monopoles of the entire EFP are scaled to ensure a net integer charge and the dipole polarizability tensor due to each LMO in the EFP region is calculated by a perturbation expression, as described in ref 36.

Free Energy. The vibrational free energy (G^{vib}) of the optimized part of the ab initio region is calculated by the partial Hessian vibrational analysis (PHVA) method.[76]

Solvation Energy. The solvation energy (ΔG_s) is calculated using the ONIOM-PCM/X approach[82] that combines IEF-PCM/ICOMP = 0 protein solvation energies with D-PCM/ICOMP = 4 solvation energies of model systems, The model system

$$\begin{aligned}\Delta G_s(\text{Protein:D-PCM/ICOMP} = 4) = {} & \\ & \Delta G_s(\text{Protein:IEF-PCM/ICOMP} = 0) + \\ & \Delta G_s(\text{Model:D-PCM/ICOMP} = 4) - \\ & \Delta G_s(\text{Model:IEF-PCM/ICOMP} = 0) \quad (3)\end{aligned}$$

consists of the side chains of Lys55 and Tyr20 (shown here for the protonated state).



UAHF spheres are used for the atoms of the entire system, and the cavitation and dispersion−repulsion energies are calculated as above. The protein solvation energies are calculated using the EFP/IEF-PCM interface developed by Bandyopadhyay, Gordon, Mennucci, and Tomasi[77,78] and extended to protein-sized systems for the calculations described here by decreasing the memory requirement and parallelizing the code.

The EFP/PCM interface is similar to an all-ab initio PCM calculation except that the electrostatic potential (**V**) of the EFP region is due to its multipole representation of the electrostatic potential. The induced surface charges influence the induced dipoles, and this contribution is iterated to self-consistency. In this study, we found several cases of divergence, presumably where surface charges are close to a polarizability tensor. Thus, the polarizability tensors are removed for the single point calculations necessary for the solvation energies.

In the current implementation, $\Delta G_{\mathrm{disp-rep}}$ [cf. eq 3] is calculated only for the ab initio and buffer region. Furthermore, surface smoothing by the generation of additional spheres[83] is prevented (by using RET = 100 in the $PCM group), because the number of added spheres never converged for the protein within the memory available.

p*K*$_a$ of Lys55. The pK_a of Lys55 computed using this approach is 11.3 pH units, which is in good agreement with the experimental value[79] of 11.1 considering the uncertainty in the experimental values is roughly ±0.1 pH units. Though this result suggests that the pK_a values of ionizable residues in proteins can be predicted with the same accuracy as that of small molecules by using QM/MM, many more cases must be investigated to test this issue. Toward this end, we addressed the two most time-consuming aspects of the QM/MM protein pK_a prediction methodology.

Three Further Methodological Improvements

Faster, Iterative Solution to the PCM Equations. The calculation of the solvation energy was by far the most time-consuming and memory intensive aspect of the pK_a prediction. For OMTKY3, roughly 10 days of CPU and 4 GB of RAM are required for the evaluation of the two solvation energies needed for a pK_a prediction using three nodes on a four-node RS/6000 44P 270 workstation. Clearly, this makes a systematic study of protein pK_a values difficult, and here we describe how we addressed this problem.

In the PCM the solute molecule is placed in the bulk solvent described as a polarizable continuum with a dielectric constant ϵ. The cavity the solute molecule occupies in the bulk solvent is defined as a set of interlocking spheres centered at atoms or atomic groups. The surface of the cavity is the boundary between the solute and solvent. In the PCM the apparent surface charge (ASC) method is used to describe the electrostatic interaction between the solute and the bulk solvent. To numerically solve the electrostatic boundary equation, the continuous charge distribution on the boundary surface is divided into a set of point charges at a finite number of boundary surface elements, called tesserae. The resulting vector of ASCs, **q**, are obtained by solving the matrix equation,

$$\mathbf{Cq} = -\mathbf{V} \quad (5)$$

where the vector **V** is the molecular electrostatic potential of the solute and **C** is a matrix that describes the interaction between the ASCs. For a small molecule with a low number of tesserae, the most efficient solution of eq 5 is matrix inversion

$$\mathbf{q} = -\mathbf{C}^{-1}\mathbf{V} \quad (6)$$

However, for larger molecules such as proteins the inversion of **C** becomes extremely memory and CPU-intensive, and an iterative solution

$$\mathbf{q}^{(n)} = -\mathbf{C}_0^{\ -1}(\mathbf{V} - \mathbf{C}_1\mathbf{q}^{(n-1)}) \quad (7)$$

proves computationally more efficient. Thus, together with Christian Pomelli, we implemented an iterative solution of the QM/EFP/PCM implementation.[84] In general, we found the use of the direct inversion of the iterative subspace (DIIS) method is needed to obtain convergence. In addition, we introduced three methodological innovations to further reduce the CPU time: (1) A looser convergence criterion for the PCM equation solution is used at early SCF steps. (2) Various multipole expansions are introduced to treat long-range electrostatic interactions. (3) Virtually all aspects of solving the IEF-PCM equations are parallelized.

The combined methodological innovations reduce the CPU and memory requirements by an order-of-magnitude.

TABLE 2: pK_a Values Computed Using Various Combinations of EFP and MM Charges[a]

method	$\Delta E^{MP2//RHF}$	ΔG_{therm}	ΔG_s	pK_a
experimental				**11.1**
14 Å EFPs + CHARMM	24.45	0.70	−24.99	10.7
14 Å EFPs + OPLSAA	24.98	0.67	−25.33	10.8
14 Å EFPs + AMBER	24.85	0.67	−25.25	10.8
14 Å EFPs + AM1	29.67	0.69	−30.62	10.4
9 Å EFPs + AMBER	25.77	0.68	−25.17	11.5
all AMBER	38.10	0.47	−40.06	9.5
all CHARMM	39.88	0.63	−42.00	9.5
all OPLSAA	39.93	0.62	−41.69	9.8
all atom-centered EFP	35.35	0.41	−30.99	14.1

[a] The pK_a values are computed as described for Lys55 in Table 1 (DPCM/ICOMP = 4) but using eqs 8−10.

Prediction of Relative pK_a Values. When implementing the iterative solution to the PCM equations, we removed an approximation introduced in the original EFP/PCM interface.[84] As a result, the predicted pK_a value of Lys55 in OMTKY3 changed from 11.3 to 10.6 pH units. Though this increases the deviation from experiment to 0.5 pH units, this error is now comparable (both in sign and magnitude) to the 0.8 unit error obtained for methylamine (Table 1). We remove this systematic error by computing the free energy (ΔG) for the following reaction

$$\mathrm{Lys55H^+ + CH_3NH_2 \leftrightarrow Lys55 + CH_3NH_3{}^+} \qquad (8)$$

via calculations of the free energy of each protonation state

$$\Delta G = [G(\mathrm{Lys55}) - G(\mathrm{Lys55H^+})] - [G(\mathrm{CH_3NH_2}) - G(\mathrm{CH_3NH_3{}^+})] \quad (9)$$

where G(X) is calculated as before [cf. eq 2]. Finally, the pK_a of Lys55H$^+$ is calculated as the pK_a shift relative to the experimental pK_a value of methylamine.

$$\mathrm{p}K_a(\mathrm{Lys55H^+}) = 10.6 + \Delta G/1.36 \qquad (10)$$

Hybrid EFP/MM Representation of the Protein. The use of the iterative PCM shifts the computational bottleneck to the construction of the protein EFP from separate ab initio calculations. We therefore investigated whether an EFP representation is needed for the entire protein or whether protein regions far from the ionizable residue could be represented by atom-centered charges from biomolecular force field such as AMBER, CHARMM, and OPLS-AA. To construct an EFP/MM representation of the protein, we had to alter our divide-and-conquer method for constructing protein EFPs as described in ref 85.

pK_a Predictions.[85] Table 2 lists the pK_a computed by a QM/buffer/EFP/MM model in which the protein environment within 14 Å is treated by an EFP and the rest of the protein is treated with either AMBER, CHARMM, or OPLSAA charges (Figure 3). The pK_a values predicted by this "four-layer" approach are all within 0.4 pH units of the experimental value, compared with a maximum error of 0.3 pH units for the all-EFP approaches. Thus, the EFP does not need to be calculated ab initio for the entire protein, leading to significant CPU-time savings. The charges from all three force fields appear of equal quality. However, using another computationally inexpensive approach such as AM1 for the >14 Å region leads to a less satisfactory pK_a of 10.4.

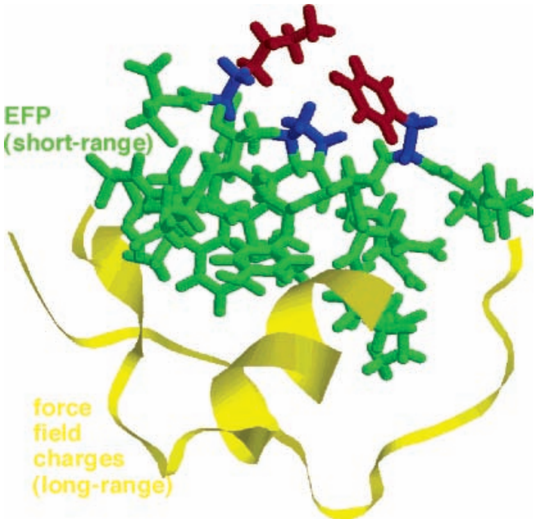


Figure 3. Ab initio/buffer/EFP/MM regions (red/blue/green/yellow) used for the computation of the pK_a of Lys55.

Decreasing the EFP region to 9 Å and using the AMBER force field for the rest of the protein does not increase the error appreciably. Still, the sole use of force field charges increases the error by as much as 1.6 pH units. It is interesting to note that the "error" (relative to the all-EFP value) in the gas-phase PA is as much as 12 kcal/mol, but that roughly 10 kcal/mol of this error is "screened" by the PCM. The error is likely due to inherent limitations in the atom-centered charge model, rather than the numerical values of the charges themselves. For example, using an all-EFP representation consisting only of atom-centered charges (but otherwise calculated as before) results in a pK_a error of 3 pH units. The better performance of the MM charges is presumably due to the underlying parametrization of the force fields against high-level ab initio calculations. In all cases discussed so far, the effect of the various representations of the protein electrostatic potential have small (≥0.62 kcal/mol) effects on the thermochemical energy contribution to the pK_a.

After the conclusion of this study it occurred to us to estimate the pK_a perturbation due to the >14 Å region, by recalculating the pK_a without this part of the EFP. Surprisingly, this resulted in a pK_a value of 10.9, indicating that the >14 Å region has negligible effect on the pK_a. This observation led us to investigate the QM-based method for pK_a prediction described next.

Minimal Model for pK_a Predictions

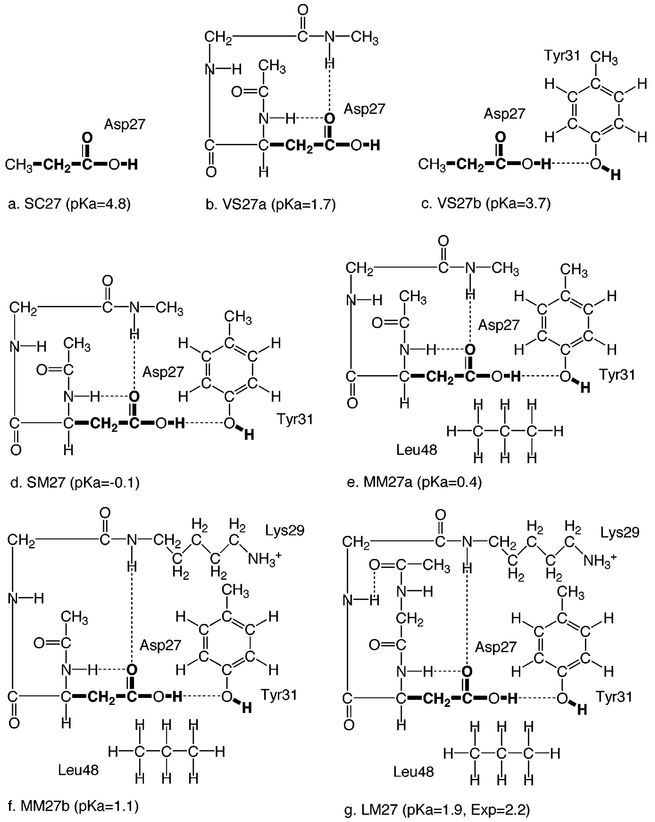
At this point we decided to reevaluate all aspects of our pK_a prediction methodology in light of the previous studies with the aim of finding the simplest possible computational model that consistently delivered accurate pK_a values for proteins.[86]

Energy Calculation. In preliminary small molecule studies (data not published) we found that the use of isodesmic reactions [e.g., eq 8] allowed us neglect the thermochemical contributions to the free energy

$$G = E_{ele} + G_{sol} \qquad (11)$$

which leads to a significant time savings because we no longer have to compute the Hessian. Similarly, the use of the isodesmic reaction allowed us to compute the solvation energy using IEF-PCM and ICOMP = 2, thus obviating the need for the ONIOM-PCM correction.

As part of this work we also discovered that the use of only 60 initial tesserae/atom resulted in lack of rotational invariance of the solvation energy, but that 240 tesserae/atom led to acceptable results.



a. SC27 (pKa=4.8) b. VS27a (pKa=1.7) c. VS27b (pKa=3.7)

d. SM27 (pKa=-0.1) e. MM27a (pKa=0.4)

f. MM27b (pKa=1.1) g. LM27 (pKa=1.9, Exp=2.2)

Figure 4. Model compounds for Asp27 of OMTKY3 and their computed pK_a values. The positions of the atoms in bold were energy minimized. Acid form I is shown.

Protein Model Construction. The approach was tested for the five Asp and Glu residues in OMTKY3. When calculating the pK_a values of carboxyl groups, we used propanoic acid as a reference compound for which the experimental values is 4.87 pH units. First a "small model" is designed that includes (1) the side-chain of the ionizable Glu or Asp residue, (2) the two amide groups next to the C$^\alpha$ of the Glu or Asp side chain, and (3) all groups that form hydrogen bonds with the carboxyl group of interest (Figure 4).

The coordinates of the atoms in each model are taken from the PDB file 1PPF.[87] Hydrogen atoms were added to the PDB structure with the WHAT IF program[88,89] at pH = 7. Several new protons were added manually to satisfy the unfilled valences where atoms were removed in constructing the small model. All of the carboxyl side chains were originally in the unprotonated form. The acid forms were obtained by adding the acidic protons to the carboxylate groups. Two or three protonation sites (i.e., conformers of the COOH group) were considered for each acid form, whereas only one base form was considered. The total free energy of the acid form is taken to be the "conformational average" of the free energies of each conformer (G_i),[90]

$$G = -RT \ln\Big[\sum_{i}^{\text{conformers}} \exp(-G_i/RT)\Big]$$
$$= G_0 - RT \ln\Big[1 + \sum_{i\neq 0}^{\text{conformers}} \exp(-\Delta G_i/RT)\Big] \qquad (12)$$

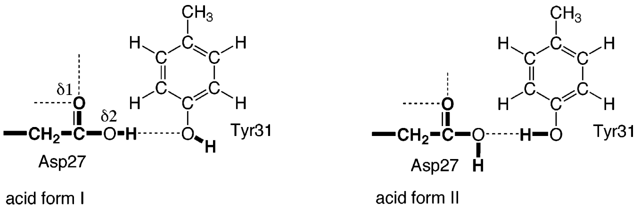
where G_0 is the lowest energy conformer and $\Delta G_i = G_i - G_0$. From the latter form of eq 12 it can be seen that only low-energy conformers ($\Delta G_i < 2RT \approx 1$ kcal/mol at 25 °C) contribute significantly to the free energy. If significant, this contribution will lower the free energy of the acid, thereby increasing the pK_a. Physically, this pK_a increase is entropic in nature, because several accessible protonation states increase the protonation probability. In our study this contribution is always less than ca. 0.2 pH units.

Because we seek the simplest possible computational model that consistently reproduces the experimental pK_a values, we optimize only a few structural parameters. The positions of the atoms in the carboxyl group (CH_2COO^- or CH_2COOH) are optimized by energy minimization, except that the Cartesian coordinates of one of the oxygens are kept fixed (except for Glu43). This allows for the carboxyl bond lengths and angles to adjust to the change in protonation state without greatly altering the overall structure. Additionally for Asp7, Glu19, and Asp27, the positions of the neighboring OH protons of Ser9, Thr17, and Tyr31, respectively, are also optimized because their positions are predicted to depend significantly on the protonation state of the carboxyl group.

Often the pK_a predicted using the small model (SM) is quite close to the experimental value, indicating that the most important intraprotein interactions are included in the SM. To analyze the interactions further, we construct several small models in which key hydrogen bonds are removed; these are called very small (VS) models (see, e.g., Figure 4). We also construct a side chain (SC) model, in which the peptide backbone atoms are replaced by a methyl group, to determine the effect of the peptide groups on the pK_a.

To determine the effect of protein groups not directly hydrogen bonded to the carboxyl group, we construct several medium models (MMs) in which side chains in the vicinity of the carboxyl group are added one at a time, without geometry re-optimization. The model in which all of the neighboring groups are included is termed the large model (LM), and the pK_a obtained using this large model is taken to be our best prediction of the experimental value. In general the pK_a values change very little on going from the SM to the LM. For this reason we have not considered models larger than the LMs.

Representative Example: Computation and Rationalization of the pK_a of Asp27. The crystal structure of OMTKY3 (1PPF) shows three possible hydrogen bonds to the carboxyl group of Asp27: Tyr31-O$^\eta$H···O$^{\delta 2}$-Asp27 (O$^\eta$-O$^{\delta 2}$ distance = 2.5 Å), Asp27-NH···O$^{\delta 1}$-Asp27 (N−O$^{\delta 1}$ distance = 2.9 Å), and Lys29-NH···O$^{\delta 1}$-Asp27 (N−O$^{\delta 1}$ distance = 2.9 Å). Accordingly, the small model of Asp27 (SM27) includes these interactions in addition to the amide group of Asn28, as shown in Figure 4d. For the acid form, two different proton positions were used: Asp27-O$^{\delta 2}$H···O$^\eta$H-Tyr31 and Asp27-HO$^{\delta 2}$···HO$^\eta$-Tyr31 (acid forms I and II, respectively).



acid form I acid form II

Upon energy minimization of SM27, acid form I is the lowest in energy, with form II 3.5 kcal/mol higher in energy. Both contributions to the free energy are included[86] and result in a pK_a of −0.1 pH units, significantly less than the experimental value of 2.2. We show below that the pK_a is increased significantly in larger models.

Removal of the Tyr31 side chain (VS27a, Figure 4b) followed by geometry reoptimization, results in a pK_a increase of 1.8 pH

TABLE 3: Comparison between the pK_a's Predicted for OMTKY3 in the Current Study and Previous Studies[a]

residue	exptl[b]	current study[c]	Forsyth[d]	Nielsen[e]	Mehler[f]	Havranek[g]
Asp7	2.5	2.4	2.9	2.7	2.9	2.1
Glu10	4.1	4.3	3.4	3.6	4.1	4.0
Glu19	3.2	2.7	3.2	2.7	3.6	3.1
Asp27	2.2	1.9	4.0	3.4	3.3	2.9
Glu43	4.8	4.5	4.3	4.3	4.4	5.6
rmsd		0.3	0.9	0.7	0.6	0.5
max. error		0.5	1.8	1.2	1.1	0.8

[a] The pK_a values are calculated using eqs 8−10 [but using propanoic acid instead of methylamine], and eq 11 instead of eq 2. [b] Reference 93. [c] The predicted pK_a's based on the large models (LMs) and B3LYP instead of MP2 are 2.2(Asp7), 4.4(Glu10), 2.2(Glu19), 2.2(Asp27), and 4.9(Glu43), with a rmsd = 0.5 and the maximum error = 1.0. [d] Reference 79. [e] Reference 11. [f] Reference 9. [g] Reference 91.

units. Comparison of the VS27a pK_a (1.7) to that of the SC27 (4.8, Figure 4a) indicates that the amide hydrogen bonds are primarily involved in lowering the pK_a relative to propanoic acid.

Larger models are constructed from SM27 by adding the Leu48 side chain (MM27a, Figure 4e), followed by the Lys29 side chain (MM27b, Figure 4f), and finally part of the Ser26 and Gly25 main chain (LM27, Figure 4g). These additions increase the pK_a by 0.5, 0.7, and 0.7 pH units, respectively, so that the pK_a of LM27 is 1.9, in good agreement with the experimental value of 2.2 pH units. All groups contain aliphatic protons that are within 3.0 Å of the carboxyl oxygens of Asp27: Leu48-C$^{\gamma 1}$H′···O$^{\delta 1}$ (2.84 Å), Leu48-C$^{\gamma 1}$H···O$^{\delta 2}$ (2.81 Å), Lys29-C$^{\beta}$H···O$^{\delta 1}$ (2.87 Å), Lys29-C$^{\delta}$H···O$^{\delta 2}$ (2.98 Å), Gly25-C$^{\alpha}$H···O$^{\delta 1}$ (2.80 Å). These rather weak interactions effectively desolvate the carboxyl group of Asp27, thereby raising the pK_a. It is especially interesting to note that the aliphatic portion of a Lys residue can increase the pK_a of a neighboring carboxyl group.

Other Carboxyl pK_a Values in OMTKY3. The QM-based pK_a prediction and rationalization methodology has also been applied to the remaining four carboxyl pK_a values on OMTKY3, and the results for the largest models are listed in Table 3, together with values predicted using a variety of classical electrostatic approaches.[9,11,79,91]

Before comparing the results, we note that the classical methods are intended to predict pK_a values with relatively little user intervention, whereas our method is used to interpret pK_a values and in the case of Asp7/Glu10 relies on the experimentally determined pK_a's to determine protonation states. However, it is clear that despite using relatively small protein models, our results are at least as good as the current literature values. Unlike the current approach, the classical methods all predict a relatively high pK_a for Asp27. In general, Schutz and Warshel[23] have noted that these methods tend to underestimate pK_a shifts.

Analyses of these pK_a values reveal that the single biggest contributor to low carboxyl pK_a values in OMTKY3 is backbone amide hydrogen bonding to the carboxyl oxygens. This observation is consistent with the study by Gunner and co-workers[92] who suggested that the electrostatic potential due to amide bonds will tend to lower pK_a values of Asp and Glu residues. Furthermore, the chemical shifts of some of these amide protons are affected by the deprotonation of the neighboring carboxyl group.[93]

Hydrogen bonds from side chain hydroxyl groups of serine and tyrosine to the carboxyl groups of Asp7 and Asp27 have effects on the pK_a values that are similar to those of amide NH hydrogen bonds. Other mutagenesis studies[94−96] have shown that hydrogen bonding to neutral and charged residues can affect the pK_a of ionizable residues by up to 1.6[94] and 2.4[96] units, respectively.

TABLE 4: List of pK_a Values Computed So Far with Our QM-Based pK_a Prediction Methodology[a]

		pK_a		
protein	residue	computed	experimental	standard
RNase H1	Asp102	2.0	< 2.0	4.0
lysozyme	Glu7	2.7	2.9	4.4
lysozyme	Asp87	2.8	2.1	4.0
cryptogein	Asp21	2.5	2.7	4.0
creatine kinase	Cys282	6.1	5.6	9.1
α1-antitrypsin	Cys232	7.5	6.9	9.1
xylanase	His149	1.2	<2.3	6.4

[a] See Table 3 for details of the pK_a calculations.

Analysis of the pK_a of Asp27 shows that neighboring hydrophophic regions can raise the pK_a by as much as 0.7 pH units. Very interestingly, the aliphatic part of the Lys29 side chain is predicted to raise the pK_a of Asp27 by 0.5 pH units. The combined effect of the hydrophobic interactions on the pK_a of Asp27 is predicted to be 2.0 pH units. The importance of hydrophobic environments in determining the pK_a has been emphasized previously, in particular by Mehler, Warshel, and Garcia-Moreno.[1,9,16,97,98]

Neighboring charged residues such as Lys (for Asp7, Glu10, and Asp27) or Arg (in the case of Glu19) are predicted to have more modest effects (≤0.5 units) than hydrogen bonding on the OMTKY3 carboxyl pK_a's. In the case of Glu19, this observation is supported by experiment:[99] the pK_a of Glu19 is increased by 0.2 and 0.4 units in R21A and T17V mutants of OMTKY3.

In general, we conclude that the prime determinants of the Asp and Glu pK_a values in OMTKY3 are local interactions within ca. 4−5 Å of the ionizable residue.

QM-Based pK_a Predictions for Other Proteins

Thus far, our QM-based pK_a prediction methodology has been used to predict four other low pK_a values of Asp and Glu residues located at the N-termini of helices in three other proteins (Table 4).

Further, we have demonstrated that the methodology performs equally well for the prediction of cysteine[100] and histidine pK_a values that are significantly shifted from the usual values. Here ethanethiol and 4-methylimidazolium are used as respective reference compounds, in place of propanoic acid [cf. eq 8]. Analyses similar to that depicted in Figure 4 indicate that amide and serine hydrogen bonds are responsible for the low pK_a values of Cys282 and Cys232, whereas desolvation is responsible for the low pK_a of His149 in xylanase. This is in agreement with our main conclusion of our OMTKY3 study that hydrogen bonding and desolvation appear to be the primary determinants of protein pK_a values.

Recent Developments

The geometry optimizations have up until recently been done in the gas phase, because we encountered numerical instabilities when optimizing using the continuum solvation model. Very recently, we have solved this problem[101] and are now routinely optimizing geometries in the aqueous phase as part of our pK_a predictions. This has also allowed us to remove the constraints imposed on the carboxyl groups mentioned above. We have also recently found that B3LYP can be used instead of MP2 with no loss of accuracy in cases where hydrogen bonds are not unusually strong (e.g., for N−O distances greater than 2.8 Å). The use of B3LYP has improved the computational efficiency of our method significantly.

Furthermore, the approach has been successfully extended to the prediction of reduction potentials of metalloproteins.[102]

Finally, our analyses of pK_a determinants have led us to propose a set of quantitative structure/pK_a relationships that form the basis for our empirical protein pK_a predictor PropK_a (http://propka.chem.uiowa.edu).[24]

Conclusions and Future Directions

We have demonstrated that ab initio quantum mechanics and hybrid ab initio-QM/MM can be used, in conjunction with a continuum treatment of the solvent, can be used to accurately predict and rationalize pK_a values of select ionizable residues in proteins. Compared to existing pK_a prediction methods, our QM-based methods require significantly more CPU time but include significantly fewer empirical parameters: (1) the experimentally determined structure, which is refined by energy minimization in the region around the ionizable residues, and (2) the atomic radii used in the calculation of the solvation energy, which are obtained for individual functional groups by fits to the solvation energy of small molecules.

Thus one use of our methodology is to analyze cases where traditional force field-based pK_a prediction method fails. In cases where our method offers correct predictions, our results can then be used to identify the source of the errors in the traditional methods, such as errors in intraprotein interaction or solvation energies, or local structural rearrangements that are typically neglected. In cases where our method fails as well, the most likely sources of error are gross errors in the experimentally determined protein structure or significant structural rearrangements upon titration. Our work so far, though still at an early stage, suggests that such cases represent exceptions and that large-scale dynamical fluctuations of the protein structure have a relatively small ($<$0.5 pH unit) effect on most protein pK_a values. Rather, our most recent work suggests that the majority of pK_a values are governed by local interactions (within roughly 7 Å of the ionizable residue) and that these interactions can be quantitatively understood by a set of simple rules. On the other hand, residues for which this is not the case will likely be the most interesting.

Acknowledgment. This work was supported by a Research Innovation Award from the Research Corporation, a type G starter grant from the Petroleum Research Fund, a grant from the National Science Foundation (MCB 0209941) to J.H.J., and a grant from the National Institutes of Health (GM 46869) to A.D.R.. H.L. gratefully acknowledges a predoctoral fellowship from the Center for Biocatalysis and Bioprocessing at the University of Iowa. The calculations were performed on IBM RS/6000 workstations and an eight-node QuantumStation obtained through NSF grants CHE 9974502 and MCB 0209941, respectively, and on computers at the Advanced Biomedical Computing Center at the National Cancer Institute at Frederick.

References and Notes

(1) Warshel, A. *Acc. Chem. Res.* **1981**, *14*, 284.

(2) Harris, T. K.; Turner, G. J. *Iubmb Life* **2002**, *53*, 85.

(3) Forsyth, W. R.; Antosiewiez, J. M.; Robertson, A. D. *Proteins-Struct. Funct. Genet.* **2002**, *48*, 388.

(4) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35*, 7819.

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

(6) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219.

(7) Ullmann, G. M.; Knapp, E. W. *Eur. Biophys. J. Biophys. Lett.* **1999**, *28*, 533.

(8) Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins-Struct. Funct. Genet.* **1993**, *15*, 252.

(9) Mehler, E. L.; Guarnieri, F. *Biophys. J.* **1999**, *77*, 3.

(10) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.

(11) Nielsen, J. E.; Vriend, G. *Proteins-Struct. Funct. Genet.* **2001**, *43*, 403.

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

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

(14) van Vlijmen, H. W. T.; Schaefer, M.; Karplus, M. *Proteins-Struct. Funct. Genet.* **1998**, *33*, 145.

(15) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721.

(16) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458.

(17) Delbuono, G. S.; Figueirido, F. E.; Levy, R. M. *Proteins-Struct. Funct. Genet.* **1994**, *20*, 85.

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

(19) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.

(20) Sandberg, L.; Edholm, O. *Proteins-Struct. Funct. Genet.* **1999**, *36*, 474.

(21) Wisz, M. S.; Hellinga, H. W. *Proteins-Struct. Funct. Genet.* **2003**, *51*, 360.

(22) Antosiewicz, J.; Briggs, J. M.; Elcock, A. H.; Gilson, M. K.; McCammon, J. A. *J. Comput. Chem.* **1996**, *17*, 1633.

(23) Schutz, C. N.; Warshel, A. *Proteins-Struct. Funct. Genet.* **2001**, *44*, 400.

(24) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins-Struct. Funct. Bioinfor.*, accepted.

(25) Buckingham, A. D.; Fowler, P. W. *J. Chem. Phys.* **1983**, *79*, 6426.

(26) Buckingham, A. D.; Fowler, P. W. *Can. J. Chem.-Rev. Can. Chim.* **1985**, *63*, 2018.

(27) Stone, A. J.; Price, S. L. *J. Phys. Chem.* **1988**, *92*, 3325.

(28) Price, S. L.; Harrison, R. J.; Guest, M. F. *J. Comput. Chem.* **1989**, *10*, 552.

(29) Sokalski, W. A.; Maruszewski, K.; Harihan, P. C.; Kaufman, J. J. *Int. J. Quantum Chem. Quantum Biol. Symp.* **1989**, *16*, 119.

(30) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 4915.

(31) Price, S. L.; Richards, N. G. J. *J. Comput.-Aided Mol. Design* **1991**, *5*, 41.

(32) Wiberg, K. B.; Rablen, P. R. *J. Comput. Chem.* **1993**, *14*, 1504.

(33) Dudek, M. J.; Ponder, J. W. *J. Comput. Chem.* **1995**, *16*, 791.

(34) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632.

(35) Kosov, D. S.; Popelier, P. L. A. *J. Chem. Phys.* **2000**, *113*, 3969.

(36) Minikis, R. M.; Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2001**, *105*, 3829.

(37) Tiraboschi, G.; Fournie-Zaluski, M. C.; Roques, B. P.; Gresh, N. *J. Comput. Chem.* **2001**, *22*, 1038.

(38) Matta, C. F.; Bader, R. F. W. *Proteins-Struct. Funct. Genet.* **2000**, *40*, 310.

(39) Sokalski, W. *Amino Acids* **1994**, *7*, 19.

(40) Dardenne, L. E.; Werneck, A. S.; Neto, M. O.; Bisch, P. M. *J. Comput. Chem.* **2001**, *22*, 689.

(41) Kedzierski, P.; Sokalski, W. A. *J. Comput. Chem.* **2001**, *22*, 1082.

(42) Gresh, N. *J. Chim. Phys. Phys.-Chim. Biol.* **1997**, *94*, 1365.

(43) Koch, U.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1701.

(44) Price, S. L.; Stone, A. J. *J. Chem. Soc., Faraday Trans.* **1992**, *88*, 1755.

(45) Lim, C.; Bashford, D.; Karplus, M. *J. Phys. Chem.* **1991**, *95*, 5610.

(46) Chen, J. L.; Noodleman, L.; Case, D. A.; Bashford, D. *J. Phys. Chem.* **1994**, *98*, 11059.

(47) Richardson, W. H.; Peng, C.; Bashford, D.; Noodleman, L.; Case, D. A. *Int. J. Quantum Chem.* **1997**, *61*, 207.

(48) Topol, I. A.; Tawa, G. J.; Burt, S. K.; Rashin, A. A. *J. Phys. Chem. A* **1997**, *101*, 10075.

(49) Topol, I. A.; Tawa, G. J.; Caldwell, R. A.; Eissenstat, M. A.; Burt, S. K. *J. Phys. Chem. A* **2000**, *104*, 9619.

(50) Topol, I. A.; Burt, S. K.; Rashin, A. A.; Erickson, J. W. *J. Phys. Chem. A* **2000**, *104*, 866.

(51) Topol, I. A.; Nemukhin, A. V.; Dobrogorskaya, Y. I.; Burt, S. K. *J. Phys. Chem. B* **2001**, *105*, 11341.

(52) Jang, Y. H.; Sowers, L. C.; Cagin, T.; Goddard, W. A. *J. Phys. Chem. A* **2001**, *105*, 274.

(53) Schuurmann, G.; Cossi, M.; Barone, V.; Tomasi, J. *J. Phys. Chem. A* **1998**, *102*, 6706.

(54) da Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **1999**, *103*, 11194.

(55) Silva, C. O.; da Silva, E. C.; Nascimento, M. A. C. *J. Phys. Chem. A* **2000**, *104*, 2402.

(56) Liptak, M. D.; Shields, G. C. *Int. J. Quantum Chem.* **2001**, *85*, 727.

(57) Liptak, M. D.; Shields, G. C. *J. Am. Chem. Soc.* **2001**, *123*, 7314.

(58) Toth, A. M.; Liptak, M. D.; Phillips, D. L.; Shields, G. C. *J. Chem. Phys.* **2001**, *114*, 4595.

(59) Li, H.; Hains, A. W.; Everts, J. E.; Robertson, A. D.; Jensen, J. H. *J. Phys. Chem. B* **2002**, *106*, 3486.

(60) Jensen, F. *Introduction to Computational Chemistry*; John Wiley & Sons, Ltd.: West Sussex, England, 1999.

(61) Barone, V.; Cossi, M.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3210.

(62) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253.

(63) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151.

(64) Floris, F. M.; Tomasi, J.; Ahuir, J. L. P. *J. Comput. Chem.* **1991**, *12*, 784.

(65) Pierotti, R. A. *Chem. Rev.* **1979**, *76*, 717.

(66) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; et al. *J. Comput. Chem.* **1993**, *14*, 1347.

(67) Cances, E.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *107*, 3032.

(68) Day, P. N.; Jensen, J. H.; Gordon, M. S.; Webb, S. P.; Stevens, W. J.; Krauss, M.; Garmer, D.; Basch, H.; Cohen, D. *J. Chem. Phys.* **1996**, *105*, 1968.

(69) Gordon, M. S.; Freitag, M. A.; Bandyopadhyay, P.; Jensen, J. H.; Kairys, V.; Stevens, W. J. *J. Phys. Chem. A* **2001**, *105*, 293.

(70) Chen, W.; Gordon, M. S. *J. Chem. Phys.* **1996**, *105*, 11081.

(71) Krauss, M.; Webb, S. P. *J. Chem. Phys.* **1997**, *107*, 5771.

(72) Day, P. N.; Pachter, R. *J. Chem. Phys.* **1997**, *107*, 2990.

(73) Merrill, G. N.; Gordon, M. S. *J. Phys. Chem. A* **1998**, *102*, 2650.

(74) Petersen, C. P.; Gordon, M. S. *J. Phys. Chem. A* **1999**, *103*, 4162.

(75) Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2000**, *104*, 6656.

(76) Li, H.; Jensen, J. H. *Theor. Chem. Acc.* **2002**, *107*, 211.

(77) Bandyopadhyay, P.; Gordon, M. S. *J. Chem. Phys.* **2000**, *113*, 1104.

(78) Bandyopadhyay, P.; Gordon, M. S.; Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **2002**, *116*, 5023.

(79) Forsyth, W. R.; Gilson, M. K.; Antosiewicz, J.; Jaren, O. R.; Robertson, A. D. *Biochemistry* **1998**, *37*, 8643.

(80) Nielsen, J. E.; McCammon, J. A. *Protein Sci.* **2003**, *12*, 313.

(81) Hoogstraten, C. G.; Choe, S.; Westler, W. M.; Markley, J. L. *Protein Sci.* **1995**, *4*, 2289.

(82) Vreven, T.; Mennucci, B.; da Silva, C. O.; Morokuma, K.; Tomasi, J. *J. Chem. Phys.* **2001**, *115*, 62.

(83) Pascualahuir, J. *J. Comput. Chem.* **1990**, *11*, 1047.

(84) Li, H.; Pomelli, C. S.; Jensen, J. H. *Theor. Chem. Acc.* **2003**, *109*, 71.

(85) Molina, P. A.; Li, H.; Jensen, J. H. *J. Comput. Chem.* **2003**, *24*, 1971.

(86) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins-Struct. Funct. Bioinform.* **2004**, *55*, 689.

(87) Bode, W.; Wei, A. Z.; Huber, R.; Meyer, E.; Travis, J.; Neumann, S. *Embo J.* **1986**, *5*, 2453.

(88) Hooft, R. W. W.; Sander, C.; Vriend, G. *Proteins-Struct. Funct. Genet.* **1996**, *26*, 363.

(89) http://www.cmbi.kun.nl/gv/servers/WIWWWI/.

(90) McQuarrie, D. A. *Statistical Thermodynamics*; University Science Books: Mill Valley, CA, 1973.

(91) Havranek, J. J.; Harbury, P. B. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11145.

(92) Gunner, M. R.; Saleh, M. A.; Cross, E.; ud-Doula, A.; Wise, M. *Biophys. J.* **2000**, *78*, 1126.

(93) Schaller, W.; Robertson, A. D. *Biochemistry* **1995**, *34*, 4714.

(94) Joshi, M. D.; Sidhu, G.; Nielsen, J. E.; Brayer, G. D.; Withers, S. G.; McIntosh, L. P. *Biochemistry* **2001**, *40*, 10115.

(95) Wang, P. F.; McLeish, M. J.; Kneen, M. M.; Lee, G.; Kenyon, G. L. *Biochemistry* **2001**, *40*, 11698.

(96) Tishmack, P. A.; Bashford, D.; Harms, E.; VanEtten, R. L. *Biochemistry* **1997**, *36*, 11984.

(97) Mehler, E. L.; Fuxreiter, M.; Garcia-Moreno, B. *Biophys. J.* **2002**, *82*, 1748.

(98) Mehler, E. L.; Fuxreiter, M.; Simon, I.; Garcia-Moreno, E. B. *Proteins-Struct. Funct. Genet.* **2002**, *48*, 283.

(99) Song, J. K.; Laskowski, M.; Qasim, M. A.; Markley, J. L. *Biochemistry* **2003**, *42*, 2847.

(100) Naor, M. M.; Jensen, J. H. *Proteins-Struct. Funct. Bioinform.* **2004**, *57*, 799.

(101) Li, H.; Jensen, J. H. *J. Comput. Chem.* **2004**, *25*, 1449.

(102) Li, H.; Webb, S. P.; Ivanic, J.; Jensen, J. H. *J. Am. Chem. Soc.* **2004**, *126*, 8010.