Accurate Prediction of Acidity Constants in Aqueous Solution via Density Functional Theory and Self-Consistent Reaction Field Methods

Jasna J. Klicić,[†] Richard A. Friesner,*[,‡] Shi-Yi Liu,[†] and Wayne C. Guida[§]

Schrödinger, Inc., 120 West 45th Street, Eighth Floor, New York, New York 10036, Chemistry Department, Columbia University, 3000 Broadway, New York, New York 10027, and Collegium of Natural Sciences, Eckerd College, 4200 54th Avenue South, St. Petersburg, Florida 33711

Received: July 5, 2001; In Final Form: November 15, 2001

We have developed a protocol for computing the acidity constant (pK_a) of organic compounds via ab initio quantum chemistry and continuum solvation methods. Density functional (DFT) calculations employing large basis sets are used to determine the gas-phase deprotonation energies. Solvation effects are treated via a self-consistent reaction field (SCRF) formalism involving accurate numerical solution of the Poisson−Boltzmann equation. Dielectric radii are parametrized for each functional group of interest to optimize solvation free energy calculations for neutral and charged species. While the intrinsic accuracy of these approaches is quite impressive (errors on the order of a few kcal/mol), it is not quite good enough to achieve the target accuracy that we have set for pK_a prediction of 0.5 pK_a units. Consequently, two further empirical parameters, scaling and additive factors, are determined for every functional group of interest by linear fitting directly to pK_a data for a training set. With this additional parametrization, an average accuracy on the order of 0.5 pK_a units is achieved. A wide range of coverage of ionizable groups is presented with special focus on chemistry of importance in pharmaceutically active compounds. In addition to obtaining data for large and diverse training sets, we have also selected a subset of known drugs for which pK_a's have been measured and made predictions for these compounds without further adjustment of parameters. The results are similar in quality to that of the training set despite the considerable size and complexity of many of these molecules, demonstrating the ability of the method to accurately handle substituent effects without explicit parametrization thereof. The method has been optimized from a computational viewpoint so that it is tractable even for relatively large pharmaceutical compounds in the 50−100 atom range.

I. Introduction

The determination of the protontation states of novel compounds in aqueous solution is a challenging and important objective of computational chemistry. At present, standard methods for pK_a prediction involve fitting of linear free energy relationships to a large empirical database. This approach can achieve high accuracy when the target functional group is well represented in the empirical database and has the advantage of requiring a minimal amount of computation time. However, it also has a number of fundamental limitations; the description of novel functional motifs, multiple functional group interactions, and electrostatic effects of the environment (e.g., for a ligand docked into a protein cavity) are likely to limit the accuracy of such empirically based calculations.

An alternative, which in principle can provide a better description of these effects, is microscopic calculations based on the underlying physical chemistry of the process. This requires quantum chemical calculation of the deprotonation event, followed by some method for evaluation of the solvation free energies of the various species. The difficulty with this approach is that the gas-phase deprotonation energy and solvation free energy difference of the protonated and deprotonated species are both large numbers which add together in the pK_a calculation with opposite signs; thus, errors of only a few percent in either number can lead to errors of a few pK_a units, which is inadequate for many of the most interesting practical applications. Finally, the computational cost of such first principles modeling is nontrivial, particularly if the level of accuracy discussed above is to be achieved.

Despite these difficulties, a number of initial efforts to compute pK_a's using high-level quantum chemical methods have been made during the past 6 years.[1,2] First, it has been established that density functional (DFT)[3] methods, in particular the hybrid B3LYP method,[4−6] is capable of achieving average errors of 1−2 kcal/mol in deprotonation energies for small molecule test cases. Second, self-consistent reaction field (SCRF)[7−9] methods have been employed in conjunction with DFT to calculate solvation energies in water. SCRF methods require parametrization of the shape of the dielectric cavity of the molecule if high accuracy is to be achieved. Results reported to date in the literature do not involve extensive parametrization and hence it is not surprising that average errors are significantly larger than 0.5 pK_a units, which is the level of accuracy one would like to achieve for problems such as structure-based drug design.

In the present paper we develop an SCRF based approach to the calculation of pK_a's which, when properly parametrized to experimental data, is capable of achieving the target accuracy specified above. We present results for a wide range of functional groups that commonly arise in pharmaceutical compounds. Parameters are fitted to small molecule data (a “training” set)

[†] Schrödinger, Inc.
[‡] Columbia University.
[§] Eckerd College.

10.1021/jp012533f CCC: $22.00
Published on Web 01/25/2002

and then tested by application to molecules obtained from the CMC database[10] of known pharmaceutical compounds. These CMC compounds are as large as 30−50 heavy atoms and consequently are large problems by quantum chemical standards. By efficient optimization of the computational protocol, combined with parallelization of the quantum chemical calculations, we are able to reduce the time to solution to an acceptable level.

The paper is organized as follows. In section II, we describe the theoretical methods used to calculate the pK_a's as well as the method and parameter optimization methodology. Section III presents the results of fitting to the molecules in the training set. The model developed in the training set is then used to predict pK_a's for a selected set of CMC compounds. Section IV, the conclusion, discusses future directions.

II. Theoretical Methodology

A. Overview. The acidity constant is a measure of a molecule's propensity to become deprotonated in aqueous solution. It is directly related to the free energy of the deprotonation reaction:

$$\mathrm{p}K_a = \frac{\Delta G_a}{2.303RT} \tag{1}$$

The deprotonation of a compound in aqueous solution can be represented as part of a thermodynamic cycle:[11]

$$\begin{tikzcd} \mathrm{AH(g)} \arrow[r, "\Delta G_g"] \arrow[d, "\Delta G_{sol}^{AH}"'] & \mathrm{A(g)}^- \arrow[d, "\Delta G_{sol}^{A^-}"'] & + & \mathrm{H(g)}^+ \arrow[d, "\Delta G_{sol}^{H^+}"'] \\ \mathrm{AH(aq)} \arrow[r, "\Delta G_a"] & \mathrm{A(aq)}^- & + & \mathrm{H(aq)}^+ \end{tikzcd}$$

One part of this cycle, ΔG_g, is the calculation of the gas-phase deprotonation energy of the molecule. Three other parts, ΔG_{sol}^{AH}, $\Delta G_{sol}^{A^-}$, and $\Delta G_{sol}^{H^+}$, are the free energy of solvation of the protonated and deprotonated form of the molecule and the proton, respectively. The fifth part of the cycle, ΔG_a, is the desired free energy of deprotonation in solution. As the sum of free energies around the cycle must add to zero, the fifth term, ΔG_a, can be calculated from the other three as

$$\Delta G_a = \Delta G_g - \Delta G_{sol}^{AH} + \Delta G_{sol}^{A^-} + \Delta G_{sol}^{H^+} \tag{2}$$

Our approach is to develop efficient methods for evaluating each part of the cycle, in some cases involving adjustable parameters.

In considering the approach to evaluating the various components of eq 2, there is one difficulty in considering each part of the cycle separately. To evaluate the energy differences required for each process, it is necessary to carry out geometry optimization of the protonated and deprotonated compounds. From a rigorous point of view, the thermodynamic cycle above is only valid if all calculations are carried out at the equilibrium geometry in solution. In many cases, replacement of the solution phase geometries with those optimized in the gas phase is a good approximation, and one that becomes better when supplemented by empirical parametrization as we describe here. However, as the molecule becomes larger and more flexible, the geometry can become a serious issue. There are several options for addressing this issue, including conformational search in solution phase using a molecular mechanics program and solution phase geometry optimization at the SCRF level. While the majority of the results below are focused on cases where this is not a serious difficulty, we do consider some examples in which, particularly for ionic as opposed to neutral species, the conformation in solution as opposed to the gas phase is a major issue. In future work this problem will be addressed more systematically.

All quantum chemical calculations described herein are carried out with the Jaguar v4.0 suite of ab initio electronic structure programs. The performance of the density functional (DFT) module in Jaguar has been presented in detail elsewhere,[12,13] and significant computational advatages were demonstrated as compared to conventional electronic structure codes. The acceleration of calculations in the present case is important to reduce the very substantial computational demands of the task at hand. The SCRF module of Jaguar has similarly been discussed in a series of papers over the past few years.[7−9] To achieve the goals of the present work, we have reoptimized parameters for the SCRF model in Jaguar for neutral solutes, fitting the results to the experimental values for free energy of transfer from the gas phase to water compiled in the literature.[14−17] For ions, however, the direct experimental data are both sparse and inaccurate. Therefore, we, in effect, parametrize ionic solvation directly to the experimental pK_a data, as is described below.

B. Gas-Phase Deprotonation Energy. The free energy of deprotonation in the gas phase is represented as

$$\Delta G_g = \Delta H_g - T\Delta S = E^{A^-} + E_{vib}^{A^-} + \frac{5}{2}RT - E^{AH} - E_{vib}^{AH} - T\Delta S^{H^+} \tag{3}$$

Here E^{AH} and E^{A^-} values are gas-phase ab initio energies of the protonated and deprotonated form of the molecule, and E_{vib}^{AH} and $E_{vib}^{A^-}$ are their respective zero-point energies. $T\Delta S^{H^+}$ is the entropic terms for H$^+$ and $^5/_2RT$ is the ideal gas approximation of the enthalpic contribution for H$^+$, assuming cancellation of equivalent terms for HA and A$^-$. We also assume that the entropic term for HA and A$^-$ will cancel out.

Work from a number of laboratories[18,19] has suggested that gas-phase deprotonation energies can be calculated via hybrid density functional theory with good accuracy as long as basis sets of sufficient quality are used. We use the B3LYP functional, which has given the best performance in tests reported in the literature, for all aspects of the calculation. Geometry optimization is carried out with the 6-31G* basis set. After the geometry optimization is converged, we use the B3LYP functional and the cc-pVTZ basis set of Dunning[20] to compute a single point energy. Diffuse functions are added at the reactive center to allow improved modeling of negative ions, with a minimal increase in CPU time. Calculations along these lines are carried out for both the neutral and ionic species, and the difference between them yields the gas-phase deprotonation energy.

To this quantity should be added the difference in zero point energies. However, our tests have shown that the difference in zero point energies between the neutral and ionic species is fairly constant as long as the functional group under study remains the same. Thus, we simply incorporate the zero point energy difference into the parametrization of each functional group, without any significant loss of accuracy. As is described below in the Results, this approach, which saves considerable CPU time, is an excellent approximation for all of the cases we have studied to date.

The DFT methods and basis sets described here are capable of yielding an average error of 1.3 kcal/mol for gas phase deprotonation energies when compared with experimental data for a series of small molecules. Here the zero point energy difference is included, since calculations do not include any

TABLE 1: Gas-Phase Deprotonation Energies in kcal/mol

molecule	ΔH_g calc	ΔH_g exp	dev
acetaldehyde	366.45	365.8	0.7
acetone	368.31	369.1	−0.8
benzene	402.78	401.7	1.1
diazomethane	373.85	373.0	0.8
4-hydroxybenzaldehyde	330.95	332.8	−1.8
4-chlorophenol	340.00	336.1	3.9
cyclopentadiene	355.25	353.9	1.4
diazirane	399.37	401.0	−1.6
ethene	409.16	409.4	−0.2
ethyne	378.50	378.0	0.5
4-fluorophenol	344.28	346.6	−2.3
hydrogen	400.98	400.4	0.6
water	389.52	390.7	−1.2
hydrogen peroxide	374.26	375.9	−1.6
hydrogen sulfide	351.19	351.1	0.1
hydrogen chloride	332.48	333.4	−0.9
hydrogen cyanide	351.03	351.4	−0.4
formic acid	343.62	345.0	−1.4
hydrogen fluoride	369.03	371.5	−2.5
dimethyl sulfide	395.79	393.2	2.6
chloromethane	396.61	396.1	0.5
acetonitrile	373.39	372.9	0.5
methyl propionate	374.20	371.9	2.3
fluoromethane	409.44	409.0	0.4
methanol	381.77	381.6	0.2
p-cresol	351.07	350.2	0.8
methanethiol	357.17	356.9	0.3
methane	419.12	416.7	2.4
4-aminophenol	354.28	352.4	1.9
ammonia	403.64	404.0	−0.4
nitromethane	355.67	356.4	−0.7
4-nitrophenol	325.42	328.7	−3.3
p-xylol	350.76	350.2	0.5
phenol	346.42	349.0	−2.6
propene	388.72	390.8	−2.1
silane	373.37	372.2	1.2
toluene	382.86	382.3	0.6
mean abs dev			1.3

parameters but are taken at their face value. Detailed results are presented in Table 1. These results are comparable to those reported by Merrill and Kass[18] for similar computational methods. While this is an impressive performance in terms of percent error given that the calculations are entirely first principles, it is not quite good enough to provide results for pK_a's at the level of precision one would like, which is around 0.5 pK_a units.

C. Solvation Free Energy of Neutral Species. We have previously described our approach for determining the solvation free energy in pure water of neutral species, which is based on the use of self-consistent reaction field (SCRF) methods involving numerical solution of the PB equation.[21,22] In early work,[23] we employed a generalized valence bond (GVB) description of the solute electronic structure. However, DFT methods provide an equally good representation of charge density at a lower computational cost. Consequently, we have reparametrized our solvation model for neutral molecules to use DFT as the electronic structure methodology. This also fits in with the use of DFT for the entire pK_a methodology.

As in ref 23, we have developed dielectric radii for various functional groups so as to fit the experimental solvation free energies for 77 small neutral solutes. We have additionally parametrized hydrogen bonding corrections, which are necessary to modify the purely electrostatic description of hydrogen bonding inherent in a dielectric continuum model, along the lines specified in ref 23. Table 2 presents predicted solvation free energies for each solute with our new model. The results are comparable in quality to those obtained with GVB methods in ref 23, and the parameters are, in fact, not very different in detail. This is an adequate level of accuracy to meet the objectives of the present study.

D. Solvation Free Energy of Ionic Species. Molecules with a net charge—whether positive or negative—naturally have a larger solvation free energy than neutral molecules. This fact follows directly from a simple Born model of solvation. Typical small polar molecules have a solvation free energy in the 5−10 kcal/mol range; in contrast, small ions are in the 50−100 kcal/mol range. This means that to achieve a 1 kcal/mol level of accuracy in prediction solvation free energies of ions, an order of magnitude greater precision in the result is required. Furthermore, because the gas phase to water free energy of transfer is so large for ions, it is extremely difficult to obtain accurate experimental numbers; error bars are typically in the 5−10 kcal/mol range and there are not a large number of values that have been obtained even at that level of precision. In contrast, pK_a's themselves can be measured quite accurately and are available for a large number of molecules.

All of this suggests that the best way to determine solvation free energies of ions is to use the thermodynamic cycle of eq 2 in conjunction with the experimental pK_a to solve for the solvation free energy of the ion. Using the methods to compute the other two legs of the cycle (gas-phase deprotonation and solvation energy of neutral molecules), we can obtain ionic solvation free energies to a reasonable level of precision.

An equivalent strategy from the standpoint of pK_a prediction is to adjust the dielectric radii of the ionic species, and the empirical correction factors, to fit experimental pK_a's. This is the approach we take here. The empirical corrections can be taken to include first shell hydrogen bonding corrections for the ionic group. These are expected to be larger for ionic molecules than for neutral ones because the gas phase hydrogen bonding energy of a molecular pair involving a charged group is typically 3−5 times larger than that for a neutral species. In the present work, these corrections, developed for each functional group, are concatenated with corrections for other transferable errors in the thermodynamic cycle.

E. Empirical Corrections. Our strategy is to develop a set of empirical parameters for a wide range of functional groups. For each functional group, a set of molecules with experimentally known pK_a's are assembled to be the “training set”. The training set is designed to include a range of substituent types so that the experimental pK_a varies over a considerable range. The methods described in the previous sections are then applied for varying values of the radius of the ionic species (all other radii being obtained from the standard list of neutral solvation parameters) and the resulting “raw” pK_a that is obtained is corrected via a simple linear fit; the final pK_a is given by the formula:

$$\mathrm{p}K_a = A(\mathrm{p}K_a^{\mathrm{raw}}) + B \qquad (4)$$

There are thus three parameters for each functional group: the radius of the ion and the constants A and B. The best values of A and B for each radius are obtained from a linear least-squares fit to the experimental training set data; then, the value of the radius that minimizes the least squares residual is chosen. van der Waals atomic radii for solvation are incorporated into the ab initio program Jaguar and are available from the authors upon request. Linear fitting parameters are listed in the Table 4.

A central difference between this approach and fully empirical pK_a prediction methods is that there are no parameters associated with substituents. It is assumed that the quantum chemical calculations are sufficiently robust to describe substituent effects

TABLE 2: Solvation Energies of Organic Molecules in kcal/mol

molecule	calc	exp	dev	molecule	calc	exp	dev
				Alkanes			
butane	2.1	2.2	-0.1	octane	3.2	2.9	0.3
ethane	1.8	1.8	0.0	pentane	2.5	2.3	0.2
heptane	2.9	2.6	0.3	propane	2.0	2.0	0.0
hexane	2.7	2.6	0.1	cyclooctane	2.1	0.9	1.2
methane	1.3	1.9	−0.6	cyclohexane	2.6	1.2	1.4
2-methylpropane	1.5	2.3	−0.8	cyclopentane	2.9	1.2	1.7
neopentane	1.7	2.5	−0.8	cyclopentene	1.6	0.6	1.0
				Alkenes			
cyclopropane	0.3	0.8	−0.5	ethene	1.0	1.3	−0.3
1,3-butadiene	0.1	0.6	−0.5	*E*-2-pentene	1.3	1.4	−0.1
cyclopentene	1.6	0.6	1.0	propene	0.8	1.3	−0.5
				Alkynes			
butenyne	−0.2	0.0	−0.2	1-pentyne	0.4	0.0	0.4
1-butyne	0.1	−0.2	0.3	propyne	−0.4	−0.3	−0.1
ethyne	−0.2	0.0	−0.2				
				Aromatics			
anthracene	−1.6	−4.2	2.6	naphthalene	−1.8	−2.4	0.6
benzene	−0.6	−0.9	0.3	*o*-xylene	−0.8	−0.9	0.1
ethylbenzene	−0.4	−0.8	0.4	toluene	−0.9	−0.8	−0.1
				Alcohols			
butanol	−4.6	−4.7	0.1	1-propanol	−4.9	−5.1	0.2
ethanol	−5.1	−5.0	−0.1	2-propanol	−5.0	−4.8	−0.2
hexanol	−4.1	−4.4	0.3	1,2-dimoxyethane	−4.1	−4.8	0.7
methanol	−5.3	−5.1	−0.2	1-methoxypropane	−1.4	−1.7	0.3
pentanol	−4.3	−4.5	0.2	2-methoxy-2-methylpropane	−1.7	−2.2	0.5
prop-2-enol	−5.1	−5.0	−0.1	2-methoxyethanol	−7.6	−6.8	−0.8
				Ethers			
dimethyl ether	−1.8	−1.9	0.1	1,4-dioxane	−5.2	−5.1	−0.1
diethyl ether	−1.4	−1.6	0.2				
				Ketones and Aldehydes			
ethanal	−3.6	−3.5	−0.1	4-methyl-2-pentanone	−3.8	−3.1	−0.7
benzaldehyde	−4.1	−4.0	−0.1	acetone	−4.4	−3.9	−0.5
butanal	−2.9	−3.2	0.3	acetophenone	−4.6	−4.6	−0.0
propanal	−3.1	−3.4	0.3	butanone	−3.6	−3.6	−0.0
2-pentanone	−3.4	−3.5	0.1	heptanone	−2.3	−2.9	0.6
3-pentanone	−3.4	−3.4	0.0				
				Carboxylic Acids			
acetic acid	−7.1	−6.7	−0.4	propanoic acid	−6.4	−6.5	0.1
butanoic acid	−6.1	−6.4	0.3				
				Amines			
butylamine	−3.9	−4.4	0.5	dimethylamine	−4.4	−4.3	−0.1
ethylamine	−4.7	−4.5	−0.2	piperazine	−8.3	−7.4	−0.9
methylamine	−4.6	−4.5	−0.1	pyrrolidine	−5.6	−5.5	−0.1
propylamine	−4.2	−4.4	0.2	morpholine	−7.6	−7.2	−0.4
azetidine	−4.3	−5.6	1.3	*N*,*N*′-dimethylpiperazine	−7.6	−7.6	−0.0
diethylamine	−3.8	−4.1	0.3	trimethylamine	−3.2	−3.2	0.0
				Nitriles			
acetonitrile	−4.7	−3.9	−0.8	propanitrile	−3.8	−3.9	0.1
				Nitro Compounds			
2-nitropropane	−3.3	−3.1	−0.2	nitroethane	−3.8	−3.7	−0.1
nitrobenzene	−3.7	−4.1	0.4	mean abs dev			0.4

TABLE 3: Acidic and Basic Functional Groups for Which Parameters Have Been Developed

acids	bases
alcohols	amines
phenols	anilines
carboxylic acids	arom. heterocycles
thiols	diazepines
sulfonamides	amidines
hydroxamic acids	guanidines
imides	pyrroles
barbituric acids	indoles
tetrazoles	

without additional parametrization. This greatly reduces the amount of parametrization required and allows extension of the calculations to novel molecular structures, perhaps containing a pattern of substituents not seen in the training set. The method also allows for environmental electrostatic effects to influence the pK_a in a natural way. Of course, more empirical information and parametrization could be input and this would undoubtedly lead to improved accuracy, at least within the training set, but would also curtail the generality of the method application. We have not explored such a strategy in this initial implemenation of the methodology.

The empirical terms are an attempt to account for both systematic contributions not explicitly calculated in the methodology described above, and for a wide range of possible errors in the thermodynamic cycle describe in eq 2. Systematic contributions include the difference in zero point energy between

TABLE 4: Linear Fit Parameters, pK_a = A(pK_a^{raw}) + B

molecule	A	B
alcohols	0.7629	−6.391
phenols	0.4713	0.631
carboxylic acids	0.4035	0.155
thiols	1.0760	−6.894
sulfonamides	0.6768	−5.556
hydroxamic acids	0.2763	3.428
imides	0.3405	1.321
barbituric acids	0.2322	2.982
tetrazoles	0.0881	3.819
primary amines	0.3009	5.110
secondary amines	0.7705	−5.305
tertiary amines	0.7043	−5.412
anilines	0.5339	−2.863
heterocycles	0.8028	−6.166
diazepines	0.7694	−7.317
amidines	1.2977	−16.970
guanidines	0.6263	−2.188
pyrroles	0.5950	−6.316
indoles	0.8424	−7.889

the neutral and ionized form of the solute if this is not explicitly calculated. Possible errors in the thermodynamic cycle are as follows:

Errors in calculated gas-phase deprotonation energies.

Residual errors in the neutral solvation calculations.

Numerical errors in obtaining solutions to the SCRF equations.

Errors in the experimental data.

First shell hydrogen bonding contributions that are not reliably computed from the continuum electrostatic model.

It is difficult to assess the relative quantitative importance of each of these terms. However, what is clear from the results is that the functional form we have used provides a good description of the overall error in the great majority of cases that we have examined. The constant term and linear term in pK_a correlate exceptionally well with the experimental data. From a physical point of view, this is not surprising. As will be shown below, some terms such as the difference in zero point energy between the neutral and ionic forms are to a very good approximation constant for a given functional group, which is more or less independent of substituent effects or the remainder of the molecule in general. In other cases, such as hydrogen bonding corrections, the term linear in pK_a is in essence measuring the charge on the ionic center. For example, in considering deprotonation of a neutral molecule to form the conjugate base, a negative ion, a lower pK_a is generated by substituents that withdraw negative charge from the deprotonated atom, thereby stabilizing the ionized form. This term is partially compensated by the less effective solvation of the ionized species due to delocalization of the charge. The main point, however, is that the magnitude of the first shell hydrogen bonding correction can be expected to depend on the size of the charge on the ionized group, and the pK_a, for reasons given above, is proportional to this quantity. Similar arguments can be made for positively charged conjugate acids.

III. Results

A. Training Set Results for Functional Groups. In our initial implementation we have included the most important acidic and basic functionalities, especially paying attention to give priority to the groups commonly encountered among medicinal compounds. The list of parametrized groups is shown in Table 3.

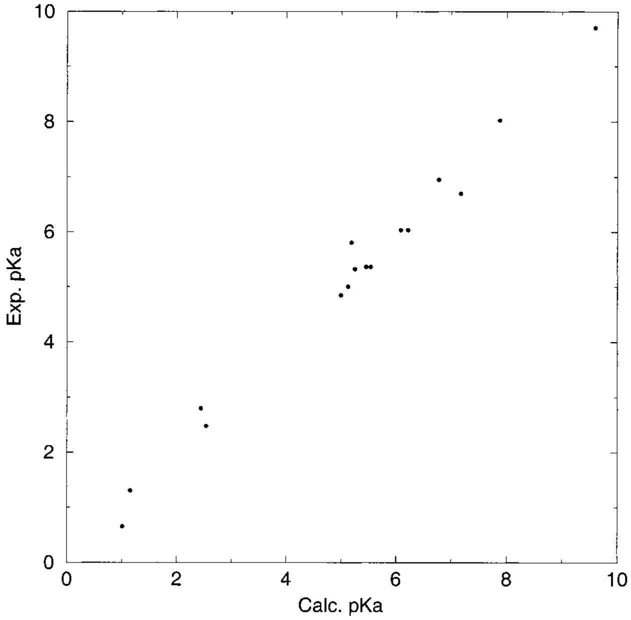

Figure 1. Experimental versus calculated pK_a values of nitrogen heterocycles.

Table 5 presents the results of the pK_a calculations for molecules in the training set as compared to the experimental data. The results are classified by functional groups.

To show a typical fit, the correlation of calculated and experimental values for aromatic nitrogen heterocycles is shown in Figure 1. It can be concluded that the performance of the model for the training set is very satisfactory.

As discussed above, we have developed a computational method where the zero point difference is included implicitly as a parameter for a given functional group rather than explicitely calculated. The advantage of this approach is a significant saving in the CPU time needed for calculations. To demontstrate the validity of this approach, we compare its performance in terms of accuracy with the method where zero point differences are calculated explicitly. The total mean average deviation for the faster method at 0.42 compares excellently with 0.41 for the latter method. Results for each functional group are presented in Table 6.

B. Tests of the Methodology. For testing we use two sets. One contains 19 aromatic nitrogen heterocycles, not dissimilar from those in the training set. The other set is more advanced and includes molecules obtained from the CMC list of medicinal compounds. There are 900 such compounds in the CMC database for which experimental pK_a values are reported. We have taken a subset of these molecules and calculated the pK_a using the method described above. Many of these molecules are quite large and complex, with multiple functional groups that interact mutually. Furthermore, none of the molecules we report below were included in the training set. Thus, the results represent a genuinely objective test of the performance of the methodology for real world problems.

Table 7 presents predicted pK_a's for nitrogen heterocycles. The average error is 0.5 pK_a units. The results for the more diverse set of molecules from the CMC database are shown in Table 8. Structures of these molecules are shown in Figure 2.

To calculate the second acidity/basicity constant, the same methodology can be applied. We have performed these simulations successfully. As an example, we can mention cysteamine from the CMC database (see Table 8). Here the thiol group is deprotonated in the presence of the protonated amino group to produce the zwitterion.

TABLE 5: Acidity Constants of Organic Acids and Bases (Mean Abs Dev 0.41)

molecule	pK_a calc	pK_a exp	dev	molecule	pK_a calc	pK_a exp	dev
Acids							
Alcohols							
methanol	16.4	15.5	0.9	2,2-dichloroethanol	12.6	12.2	0.3
ethanol	16.0	15.9	0.1	2,2,2-trichloroethanol	11.6	12.4	−0.8
propanol	16.0	16.2	−0.2	2,2,2-trifluoroethanol	15.6	15.4	0.2
2−propanol	15.9	17.1	−1.2	1,2-ethanediol	13.4	13.6	−0.1
2-butanol	16.9	17.6	−0.7	1,2-propanediol	15.4	14.9	0.5
tert-butyl alcohol	16.2	19.2	−3.0	1,3-propanediol	16.4	15.1	1.3
allyl alcohol	15.3	15.5	−0.2	1,4-butanediol	16.4	15.1	1.3
propargyl alcohol	15.0	14.3	0.7	mean abs dev			0.8
2-chloroethanol	13.8	12.9	0.9				
Phenols							
phenol	9.8	10.0	−0.2	4-methylphenol	10.6	10.5	0.2
4-aminophenol	9.3	9.4	−0.1	4-nitrophenol	7.3	7.2	0.2
4-chlorophnol	9.6	9.9	−0.3	*p*-xylol	10.4	10.3	0.0
4-fluorophenol	10.4	10.2	0.2	4-hydroxybenzaldehyde	7.6	7.6	0.0
4-methoxyphenol	10.3	10.3	0.0	mean abs dev			0.1
Carboxylic Acids							
cia-1,2-cyclopropanedicarboxylic acid	4.3	3.6	0.7	acetic acid	3.7	4.8	−1.1
trans-1,2-cyclopropanedicarboxylic acid	3.9	3.8	0.1	acrylic acid	3.8	4.2	−0.5
cis-2-chlorobut-2-enecarboxylic acid	3.3	2.8	0.5	benzoic acid	3.9	4.2	−0.3
trans-2-chlorobut-2-enecarboxylic acid	3.0	3.2	−0.2	butanoic acid	4.2	4.8	−0.6
2-chlorobut-3-enecarboxylic acid	2.7	2.5	0.2	t cinnamic acid	4.3	4.4	−0.1
2-chloropropanecarboxylic acid	3.0	2.9	0.1	formic acid	2.9	3.8	−0.8
2,2-dimethylpropanoic acid	4.3	5.0	−0.8	glycolic acid	3.4	3.8	−0.5
2-furanecarboxylic acid	3.3	3.2	0.1	glyoxylic acid	1.6	2.3	−0.7
cis-2-methylcyclopropanecarboxylic acid	4.1	5.0	−0.9	malic acid	2.7	3.5	−0.8
trans-2-methylcyclopropanecarboxylic acid	4.4	5.0	−0.6	malonic acid	3.4	2.9	−.6
2-methylpropanecarboxylic acid	4.5	4.6	−0.1	oxalic acid	2.0	1.2	0.8
cis-3-chlorobut-2-enecarboxylic acid	3.9	4.1	−0.2	pentafluoropropanoic acid	0.5	-0.4	0.9
trans-3-chlorobut-2-enecarboxylic acid	3.5	3.9	−0.5	propanoic acid	4.1	4.9	−0.8
3-chloropropanecarboxylic acid	4.3	4.1	0.2	propargylic acid	2.7	1.9	0.8
cis-3-chloropropenecarboxylic acid	3.9	3.5	0.4	succinic acid	4.1	4.2	−0.2
trans-3-chloropropenecarboxylic acid	3.6	3.8	−0.2	*dl*-tartaric acid	3.2	3.0	0.2
3-chloropropynecarboxylic acid	2.9	1.9	1.0	*meso*-tartaric acid	2.4	3.2	−0.8
3-nitro-2-propanecarboxoxylic acid	4.5	2.6	1.9	tartonic acid	2.4	2.4	0.0
3-oxopropanecarboxylic acid	5.3	3.6	1.7	trifluoroacetic acid	0.4	0.2	0.2
cis-4-chlorobut-3-enecarboxylic acid	4.4	4.1	0.3	mean abs dev			0.5
trans-4-chlorobut-3-enecarboxylic acid	3.9	4.1	−0.2				
Thiols							
methanethiol	10.0	10.3	−0.3	1,2-ethanedithiol	9.2	9.1	0.2
ethanethiol	10.8	10.6	0.2	thiophenol	6.6	6.6	0.0
2-merkaptoethanol	9.4	9.4	−0.0	mean abs dev			0.2
Sulfonamides							
N-chlorotolylsulfonamide	4.3	4.5	−0.2	sulfadiazine	7.0	6.5	0.5
dichlorphenamide	6.5	7.4	−0.9	sulfadimethoxine	7.2	6.0	1.2
mafenide	9.4	8.5	0.9	sulfamethazine	7.7	7.4	0.3
methanesulfonamide	10.1	10.5	−0.4	sulfanylamide	10.4	10.4	−0.1
nimesulide	6.3	5.9	0.4	sulfapyridine	7.8	8.4	−0.6
quinethazone	9.1	9.3	−0.2	sulfaquinoxaline	6.4	5.5	0.9
saccharin	3.0	1.6	1.4	sulthiame	9.1	10.0	−0.9
sulfamethizole	3.2	5.4	−2.3	xipamide	9.3	10.0	-0.7
sulfaperin	7.2	6.8	0.5	mean abs dev			0.7
sulfacetamide	5.6	5.4	0.2				
Hydroxamic Acids							
formohydroxamic acid	8.0	8.7	−0.6	3-nitrobenzohydroxamic acid	8.2	8.4	−0.2
acetohydroxamic acid	8.5	8.7	−0.2	4-aminobenzohydroxamic acid	8.8	9.4	−0.6
benzohydroxamic acid	8.5	8.8	−0.3	4-chlorobenzohydroxamic acid	8.4	8.7	−0.3
salicylhydroxamic acid	8.4	7.5	1.0	4-flurobenzohydroxamic acid	8.4	8.8	−0.4
2-aminobenzohydroxamic acid	9.0	9.0	−0.0	4-nitrobenzohydroxamic acid	8.2	8.3	−0.1
2-chlorobenzohydroxamic acid	8.3	7.8	0.5	4-hydroxybenzohydroxamic acid	8.6	8.9	−0.3
2-fluorobenzohydroxamic acid	8.2	8.0	0.2	mean abs dev			0.4
2-nitrobenzohydroxamic acid	8.5	7.0	1.4				
Imides							
fluorouracil	8.6	8.0	0.6	dimethadione	7.6	6.1	1.5
methylthiouracil	7.9	8.2	−0.3	phthalimide	8.8	9.9	−1.1
phenytoin	8.0	8.3	−0.3	succinimide	8.7	9.6	−0.9
3,3-methylphenylglutarimide	10.2	9.2	1.0	mean abs dev			0.8
3,3-dimethylsuccinimide	8.9	9.5	−0.6				

TABLE 5 (Continued)

molecule	pK_a calc	pK_a exp	dev	molecule	pK_a calc	pK_a exp	dev
Barbituric Acids							
5,5-methylphenylbarbituric acid	7.5	7.4	0.1	5,5-dimethylbarbituric acid	8.1	8.0	0.1
1,5,5-trimethylbarbituric acid	8.3	8.3	−0.0	1,5-dimethyl-5-phenylbarbituric acid	7.6	7.8	−0.2
hexobarbital	8.2	8.2	−0.0	mean abs dev			0.1
Tetrazoles							
5-cyclopropyltetrazole	4.9	5.4	−0.5	5-phenyltetrazole	5.0	3.5	1.5
5-methyltetrazole	4.8	5.6	−0.8	tetrazole	4.8	4.9	−0.1
5-hydroxytetrazole	5.0	5.4	−0.4	mean abs dev			0.6
5-phenoxytetrazole	4.6	4.4	0.2				
Bases							
Primary Amines							
methylamine	10.5	10.2	0.3	2-aminoethanol	9.8	9.2	0.6
ethylamine	11.0	10.6	0.3	1,2-ethanediamine	10.1	10.7	−0.6
propylamine	10.7	10.6	0.1	1,3-propanediamine	10.4	10.9	−0.5
tert-butylamine	10.5	10.7	−0.2	mean abs dev			0.4
Secondary Amines							
dimethylamine	10.9	10.7	0.2	piperidine	11.1	11.1	−0.0
diethylamine	11.1	11.0	0.0	morpholine	9.5	8.5	1.0
azetidine	11.3	11.3	−0.0	2,5-diazahexane	9.4	10.4	−1.0
pyrrolidine	11.1	11.3	−0.1	mean abs dev			0.3
Tertiary Amines							
trimethylamine	10.1	9.8	0.3	dimethylcyclohexylamine	10.6	10.7	−0.1
triethylamine	10.6	11.0	−0.4	dimethylbenzylamine	8.9	9.0	−0.1
tripropylamine	9.2	10.7	−1.4	diethylbenzylamine	9.2	9.5	−0.2
1-methylpiperidine	10.4	10.2	0.2	hexamethylenetetramine	6.5	5.3	1.3
triallylamine	7.1	8.3	−1.3	DABCO	9.6	8.2	1.4
1-allylpiperidine	9.9	9.7	0.3	mean abs dev			0.6
Anilines							
aniline	4.7	4.6	0.1	4-nitroaniline	1.1	1.0	0.1
4-chloroaniline	4.0	4.0	0.1	p-toluidine	4.6	5.1	−0.5
4-methoxyaniline	5.5	5.2	0.3	mean abs dev			0.2
Heterocycles							
2-aminopyridine	7.2	6.7	0.5	melamine	5.1	5.0	0.1
2-aminothiazole	5.5	5.4	0.2	pyrazine	1.0	0.7	0.4
2-methylimidazole	7.9	8.0	−0.1	pyrazole	2.5	2.5	0.1
3-aminopyridine	6.1	6.0	0.0	pyridine	5.2	5.3	−0.1
4-aminopyridine	9.6	9.7	−0.1	pyrimidine	1.1	1.3	−0.2
4-methylpyridine	6.2	6.0	0.2	quinoline	5.0	4.8	0.1
benzimidazole	5.2	5.8	−0.6	thiazole	2.4	2.8	−0.4
imidazole	6.8	7.0	−0.2	mean abs dev			0.2
isoquinoline	5.4	5.4	0.1				
Amidines							
hydroxyimidazo[2,3-*a*]isoindole	9.1	8.6	0.5	tolazoline	10.6	10.3	0.3
imidazo[2,3-*b*]thioxazole	8.1	8.0	0.1	mean abs dev			0.5
tetrahydrozoline	9.6	10.5	−0.9				
Benzodiazepines							
1,3-dihydro-1-methyl-5-phenyl-1,4-benzodiazepin-2-one	3.8	3.3	0.5	1,3-dihydro-5-phenyl-1,4-benzodiazepin-2-one	4.0	3.5	0.5
1,3-dihydro-3-hydroxy-5-phenyl-1,4-benzodiazepin-2-one	1.9	1.7	0.2	2,3-dihydro-1-methyl-5-phenyl-1,4-benzodiazepine	6.1	6.2	−0.1
1,3-dihydro-3-hydroxy-1-methyl-5-phenyl-1,4-benzodiazepin-2-one	1.4	1.6	−0.2	3-hydro-2-methylamine-4-oxy-5-phenyl-1,4-benzodiazepine	3.9	4.8	−0.9
Guanidines							
clonidine	8.2	8.1	0.1	methylguanidine	13.4	13.4	0.0
debrisoquin	13.0	11.9	1.1	mean abs dev			0.6
guanidine	12.5	13.8	−1.3				
Pyrroles (C-2 Protonation)							
pyrrole	−4.1	−3.8	−0.3	3-methylpyrrole	−0.9	−1.0	0.1
1-methylpyrrole	−2.3	−2.9	0.6	mean abs dev			0.4
2-methylpyrrole	−0.7	−0.2	−0.5				
Indoles (C-3 Protonation)							
indole	−3.7	−3.6	−0.1	3-methylindole	−4.6	−4.6	−0.0
1-methylindole	−2.0	−2.3	0.3	mean abs dev			0.1
2-methylindole	−0.4	−0.3	−0.1				

The CPU time required for the completion of the pK_a computational cycle depends on the molecule being studied. The larger systems naturally require more time. However, a molecule's flexibility plays a role too. If the molecule's

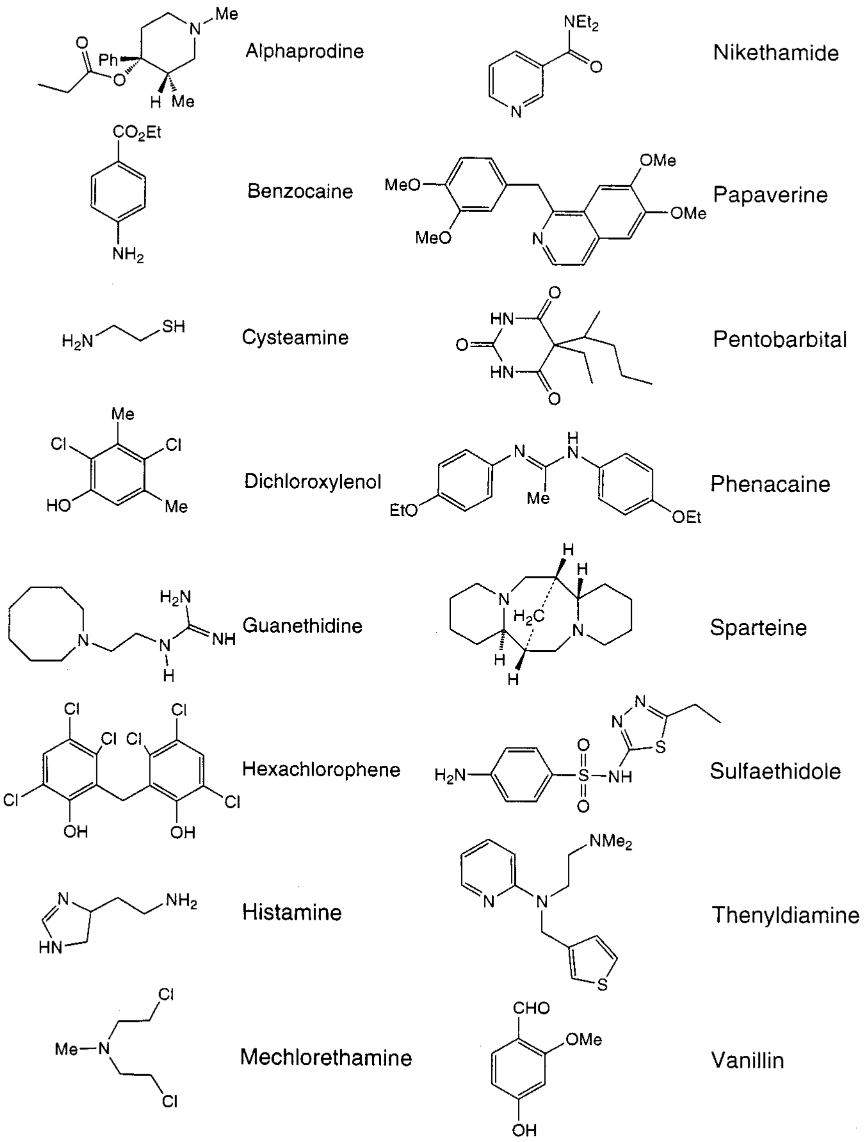

Figure 2. Druglike molecules from the CMC database.

TABLE 6: Mean Absolute Deviations of pK_a Constants for Methods with ZPE Difference Parametrized and Calculated

funct group	ZPE param	ZPE calc
alcohols	0.75	0.78
phenols	0.18	0.18
carboxylic acids	0.44	0.43
thiols	0.14	0.10
hydroxamic acids	0.49	0.40
imides	0.61	0.59
barbituric acids	0.17	0.17
tetrazoles	0.46	0.46
primary amines	0.38	0.39
secondary amines	0.37	0.38
tertiary amine	0.59	0.66
anilines	0.08	0.10
heterocycles	0.41	0.41
amidines	0.21	0.23
benzodiazepines	0.30	0.31
pyrroles	0.17	0.12
indoles	0.16	0.14
total	0.42	0.41

TABLE 7: Acidity Constants of Aromatic Heterocycles

molecule	pK_a calc	pK_a exp	dev
1,2,3-triazole	1.7	1.2	0.5
1,2,4-triazole	2.2	2.5	−0.2
1,2,5-thiadiazole	−5.3	−4.9	−0.4
benz-[3,4]-isothiazole	0.4	−0.1	0.4
benz-[3,4]-isoxazole	−2.1	−2.2	0.1
benz-[4,5]-isoxazole	−3.9	−4.7	0.8
benzoxazole	−1.5	−2.2	0.7
benzpyrazole	1.0	1.3	−0.3
benzthiazole	1.1	1.2	−0.1
isothiazole	−0.3	−0.5	0.2
isoxazole	−3.3	−3.0	−0.3
N-methyl-1,2,3-triazole	2.0	1.2	0.7
N-methyl-1,2,4-triazole	2.2	3.2	−1.0
N-methylbenz-[3,4]-pyrazole	2.4	2.0	0.4
N-methylbenzimidazole	5.5	5.5	0.0
N-methylbenzpyrazole	0.8	0.4	0.4
N-methylimidazole	7.3	7.3	−0.0
N-methylpyrazole	2.5	2.1	0.5
oxazole	0.2	2.5	−2.3
mean abs dev			0.5

conformation is relatively rigid, it will take fewer geometry optimization steps to reach the minimum energy geometry. Some representative timings are presented in Table 9. A number of basis functions for 6-31G* and cc-pVTZ(-f)[+] basis sets are shown to give an idea about the molecular size. The timing study was run on the Compaq Alpha Server DS20 computer.

For calculating pK_a constants of very large systems we recommend the following procedure: given the fact that parts

TABLE 8: Acidity Constants of Medicinal Molecules from the CMC Database

molecule	pK_a calc	pK_a exp	dev
alphaprodine	8.2	8.7	−0.5
benzocaine	2.3	2.5	−0.2
cysteamine	11.0	10.5	0.5
dichloroxylenol	8.5	8.3	0.2
guanethidine	11.9	11.4	0.5
hexachlorophene	6.1	5.7	0.4
histamine	10.0	9.7	0.3
mechlorethamine	6.1	6.4	−0.4
nikethamide	3.4	3.5	−0.1
papaverine	6.9	6.4	0.5
pentobarbital	8.0	8.0	0.0
phenacaine	9.1	9.3	−0.2
sparteine	10.6	12.0	−1.4
sulfaethidole	3.5	5.6	−2.1
thenyldiamine	7.1	8.9	−1.8
vanillin	7.8	7.4	0.4
mean abs dev			0.6

TABLE 9: CPU Time in Minutes on Compaq AlphaServer DS20

	6-31G*	cc-pVTZ(-f)[+]	CPU
molecule	bfn	bfn	time
methanol	38	91	3.3
formic acid	49	105	4.4
ethylamine	61	150	6.7
azetidine	76	173	15.5
acetohydroxamic acid	85	169	28.0
methylguanidine	91	196	15.9
pyrimidine	100	192	15.9
phenol	117	224	23.7
trans-2-methylcyclopropane-carboxylic acid	121	251	37.8
5-cyclopropyltetrazole	132	247	75.6
p-toluidine	140	283	68.5
methylthiouracil	151	274	60.4
tripropylamine	194	437	140.7
tolazoline	206	402	177.9
3-nitrobenzohydroxamic acid	207	362	236.6
debrisoquin	223	434	154.1
sulfacetamide	234	425	331.2
tetrahydrozoline	259	507	265.9
sulfaperin	298	535	305.1
phenytoin	309	554	571.7

of a molecule that are chemically and spatially distant from the group undergoing deprotonation or protonation exert very little influence on the energetics of this reaction, it is a very good approximation to construct a smaller chemical systems retaining all the key structural features to model in the pK_a calculation. Naturally, one should be aware of the possible relay effect between various polar groups in the molecule when building such a model system. In any case, it is wise to employ good chemical intuition and knowledge of the chemical system that is being studied.

IV. Conclusion and Future Directions

We have developed a methodology for calculating acidity constants of organic acids and bases in water based on the ab initio simulations. This work required selecting the right ab initio methods and basis sets that would give accurate enough results for practical applications and still enable users to complete the whole computational cycle in the reasonable amount of the CPU time, even for larger organic systems. Besides deciding on the optimal computational methods, the methodology development also required adjustment of the solvation radii for the Poisson−Boltzmann method based solvation calculations, as well as the development of the empirical fitting parameters that ensure an agreement of predicted pK_a constants with the experimental data in the absolute sense and compensate for the lack of hydrogen bonding treatment in the solvation method. The average absolute error in the training set of about 200 molecules is 0.4 pK_a units. Testing the method on a diverse set of medicinal compounds gave an average error less than one pK_a unit, which was our goal. The test results show that the method can successfully be applied in predicting the acidity constants of organic acids and bases in water.

In the future work we would like to extend this methodology to predict the acidity of a functional group affected by its chemical surrounding. An example of such a system is a titratable amino acid residue burried in a protein. Being able to estimate pK_a values of these groups would be a great success that would open doors to understanding many mechanisms of enzymatic reactions.

Acknowledgment. This work was supported in part by a grant to R.A.F. from the National Institutes of Health, Institute of General Medical Sciences (GM-40526). We also express our gratitude to Dr. Thomas A. Halgren for many fruitful discussions on the subject.

References and Notes

(1) Kallies, B.; Mitzner, R. *J. Phys. Chem. B* **1997**, *101*, 2959.

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

(3) Andzelm, J. In *Density Functional Methods in Chemistry*; Labanowski, J. K., Andzelm, J. W., Eds.; Springer-Verlag: New York, 1991; p 155.

(4) Becke, A. D. *J. Chem. Phys.* **1992**, *96*, 2155.

(5) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.

(6) Becke, A. D. *Int. J. Quantum Chem.: Quantum Chem. Symp.* **1989**, *23*, 599.

(7) Rashin, A. A.; Young, L.; Topol, I. A. *Biophys. Chem.* **1994**, *51*, 359.

(8) Bachs, M.; Luque, F. J.; Orozco, M. *Comput. Chem.* **1994**, *15*, 446.

(9) Fortuneli, A.; Tomasi, J. *Chem. Phys. Lett.* **1994**, *34*, 231.

(10) CMC-3D.; version 98.1.; MDL Infomation Systems, Inc.: 14600 Catalina St., San Leandro, CA 94577.

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

(12) Friesner, R. A.; Murphy, R. B.; Beachy, M. D.; Ringnalda, M. N.; Pollard, W. T.; Dunietz, B. D.; Cao, Y. *J. Phys. Chem. A* **1999**, *103*, 1913.

(13) Murphy, R. B.; Cao, Y.; Beachy, M. D.; Ringnalda, M. N.; Friesner, R. A. *J. Chem. Phys.* **2000**, *112*, 10131.

(14) Cramer, C. J.; Truhlar, D. G. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 629.

(15) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgage, C. C. *Biochemistry* **1981**, *20*, 849.

(16) Wolfenden, R. *Biochemistry* **1978**, *17*, 201.

(17) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *Solution Chem.* **1981**, *563*, 10.

(18) Merrill, G. N.; Kass, S. R. *J. Phys. Chem.* **1996**, *100*, 17465.

(19) Smith, B. J.; Radom, L. *Chem. Phys. Lett.* **1995**, *245*, 123.

(20) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.

(21) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1570.

(22) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591.

(23) Marten, B.; Kim, K.; Cortis, C.; Friesner, R. A.; Murphy, R. B.; Ringnalda, M. N.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775.