The Determinants of pK_as in Proteins†

Jan Antosiewicz‡ and J. Andrew McCammon

Departments of Chemistry and Biochemistry, and Pharmacology, University of California at San Diego, La Jolla, California 92093-0365

Michael K. Gilson*

Center for Advanced Research in Biotechnology, National Institute of Standards and Technology, 9600 Gudelsky Drive, Rockville, Maryland 20850-3479

Received January 22, 1996; Revised Manuscript Received April 11, 1996⊗

ABSTRACT: Although validation studies show that theoretical models for predicting the pK_as of ionizable groups in proteins are increasingly accurate, a number of important questions remain: (1) What factors limit the accuracy of current models? (2) How can conformational flexibility of proteins best be accounted for? (3) Will use of solution structures in the calculations, rather than crystal structures, improve the accuracy of the computed pK_as? and (4) Why does accurate prediction of protein pK_as seem to require that a high dielectric constant be assigned to the protein interior? This paper addresses these and related issues. Among the conclusions are the following: (1) computed pK_as averaged over NMR structure sets are more accurate than those based upon single crystal structures; (2) use of atomic parameters optimized to reproduce hydration energies of small molecules improves agreement with experiment when a low protein dielectric constant is assumed; (3) despite use of NMR structures and optimized atomic parameters, pK_as computed with a protein dielectric constant of 20 are more accurate than those computed with a low protein dielectric constant; (4) the pK_a shifts in ribonuclease A that result from phosphate binding are reproduced reasonably well by calculations; (5) the substantial pK_a shifts observed in turkey ovomucoid third domain result largely from interactions among ionized groups; and (6) both experimental data and calculations indicate that proteins tend to lower the pK_as of Asp side chains but have little overall effect upon the pK_as of other ionizable groups.

Predictive models for protonation equilibria in proteins are needed for a variety of practical problems. For example, it will be difficult to predict the affinities of proteins for ligands without knowledge of ionization states for groups in or near the binding site. The problem is made more complex because ionization states sometimes change when a ligand binds to a protein [see, for example Yamazaki et al. (1994) and Wlodek et al. (1995)]. Thus, accurate methods of accounting for protonation equilibria are likely to be helpful in structure-based drug design. More generally, detailed simulations of proteins, such as MD and Monte Carlo simulations, customarily require that the charge state of every ionizable group be specified *a priori*. Use of an inappropriate charge state for a critical residue, such as a catalytic residue in an enzyme active site, may well cause a simulation to yield unrealistic results.

The formulation of realistic and predictive models for protonation equilibria in proteins represents a fundamental challenge in physical chemistry. Published theoretical treatments apparently begin with the smeared-charge model of Linderstrom-Lang (1924), which relies upon the concept that pK_a shifts result primarily from electrostatic interactions. Published in 1924, this model predates by over three decades the availability of protein structures at atomic resolution (Kendrew, 1963) and the precise and efficient measurement of protein pK_as by NMR[1] spectroscopy (Mandel, 1964, 1965). These advances, combined with dramatic progress in computer hardware and numerical methods, now permit the development and testing of increasingly predictive models for protonation equilibria in proteins. [See, for example, Warshel (1981), Gilson and Honig (1987), Bashford and Karplus (1990), Langsetmo et al. (1991), Beroza et al. (1991, 1995), Takahashi et al. (1992), Oberoi and Allewell (1993), Yang et al. (1993), Yang and Honig (1993), Antosiewicz et al. (1994), and You and Bashford (1995).] However, the concept that the work of ionizing a titratable group in a protein is modulated primarily by electrostatic interactions still underlies most, if not all, current models.

In particular, a number of laboratories have used detailed solutions of the linearized PB equation for proteins in solution to estimate the energetics of ionization processes (Gilson & Honig, 1987, 1988b; Bashford & Karplus, 1990; Beroza et al., 1991; Bashford & Gerwert, 1992; Yang et al., 1993; Yang & Honig, 1993; Oberoi & Allewell, 1993; Antosiewicz et al., 1994; You & Bashford, 1995). These studies typically yield values of the pK_as of ionizable groups,

† Supported by the National Institute of Standards and Technology, the National Science Foundation, the National Institutes of Health, and the National Science Foundation Supercomputer Centers MetaCenter Program. J.A. was supported in part by the Polish Ministry of National Education (BST-501).

* Corresponding author. Voice: (301) 738-6217. FAX: (301) 738-6255. E-mail: gilson@indigo14.carb.nist.gov.

‡ Current address: Department of Biophysics, University of Warsaw, 02-089, Warsaw, Poland.

⊗ Abstract published in *Advance ACS Abstracts,* June 1, 1996.

[1] Abbreviations: MD, molecular dynamics; NMR, nuclear magnetic resonance; PB, Poisson−Boltzmann; HEWL, hen egg white lysozyme; BPTI, bovine pancreatic trypsin inhibitor; RNase A, ribonuclease A; OMTKY3, turkey ovomucoid third domain; RMSD, root mean square deviation.

S0006-2960(96)00156-0 CCC: $12.00

which may then be compared with experiment. Such comparisons test the validity of the physical model and the parameters employed. They also provide information on the accuracy that can be expected when the models are applied to proteins outside of the test set.

Although certain components of the PB model, such as the ionic strength and the dielectric constant of the solvent, are fixed by the experimental conditions to be simulated, other components are less certain. These include the atomic charges and radii used in the electrostatics calculations; the conformation or conformational distribution of the protein; and the dielectric constant of the protein. The present study makes use of the substantial body of structural and pK_a data for the proteins HEWL, RNase A, BPTI, and OMTKY3, to examine the influence of the dielectric constant of the protein, of atomic parameters, and of protein conformation, upon the accuracy of the calculations. This study aims to establish models that are increasingly accurate and thus, presumably, increasingly realistic. It is based upon the consistent application of a set of models to the four proteins, permitting cumulative comparisons of computed pK_as with measurements.

However, closer examination of some of these systems, and of the experimental data themselves, is also of considerable interest. Accordingly, an additional study of the influence of bound phosphate upon the pK_as of histidine side chains in RNase A is provided, and the basis for the pK_a shifts of carboxylic acids in OMTKY3 is analyzed. In addition, observations resulting from a cumulative analysis of 63 measured and calculated pK_as are presented.

The next three sections provide more detailed introductions to the studies outlined above.

A. Cumulative Studies of Accuracy

1. Intrinsic Dielectric Constant of a Protein. Calculations on folded proteins yield dielectric constants in the range 2−5 (Gilson & Honig, 1986; Simonson et al., 1991), except when ionized groups are considered to contribute to the fluctuations in dipole moment of the molecule (King et al., 1991; Smith et al., 1993), as recently emphasized by Simonson and Perahia (1995). It has been argued that the lower values of 2−4 are the most appropriate for use in PB calculations, because the contribution of ionizable groups to the fluctuations in dipole moment should not be included in the dielectric constant (Antosiewicz et al., 1994). It is therefore somewhat surprising that pK_as computed with an assumed protein dielectric constant of 20 agree with measured pK_as substantially better than those computed with a protein dielectric constant of 4 (Antosiewicz et al., 1994). The reason for this result is not known. Although it may be that the dielectric constant of a folded protein actually is near 20, it makes sense to consider other explanations (Antosiewicz et al., 1994). One possibility is that the charges and radii that are typically assigned to the atoms of the protein are not optimally parameterized. Another is that crystallization of the proteins for determination of their structures leads to conformational changes, and that if the calculations were carried out on conformations more appropriate to the solution conditions under which the pK_a measurements were made, accurate pK_as would be obtained with a protein dielectric constant of 4. Because pK_as computed with a protein dielectric constant of 4 depend strongly upon conformation (Bashford & Karplus, 1990; Bashford & Gerwert, 1992; Bashford et al., 1993; Yang & Honig, 1993), it may be appropriate to compute pK_as as averages over a range of conformations. The present paper considers several of these possibilities by assessing the accuracy of pK_as computed with a low protein dielectric constant, in combination with two different sets of atomic parameters, and two different treatments of protein conformation. For comparison, otherwise identical calculations carried out with a protein dielectric constant of 20.

2. Atomic Charges and Radii. The charges assigned to the neutral and ionized forms of titratable groups, as well as the charges assigned to nontitrating groups, are important parameters in electrostatic models of ionization. The radii assigned to atoms are also important, for they determine the position of the boundary between the low-dielectric interior of the molecule and the high-dielectric solvent, and thus the strength of the electrostatic interaction between solute and solvent (Gilson & Honig, 1988a). Various sets of atomic charges and radii have been used in calculations of pK_as with the PB model. Most have been based upon parameters associated with the molecular simulation packages CHARMm (Brooks et al., 1983; Bashford & Karplus, 1990; Beroza et al., 1991; Bashford & Gerwert, 1992; Oberoi & Allewell, 1993; Antosiewicz et al., 1994) or DISCOVER (Biosym Technologies, San Diego, CA, 1993; Hagler et al., 1974; Beroza & Fredkin, 1996; Yang et al., 1993; Yang & Honig, 1993; Beroza et al., 1995). One study considers five different parameter sets (Bashford et al., 1993), but it is difficult to determine which parameters are best.

Two new parameter sets have recently been developed that are specifically designed for use with the PB model of electrostatics (Sitkoff et al., 1994; Schmidt & Fine, 1994). The parameters are optimized to reproduce measured solvation energies of small molecules, when used with a solvation model based upon the PB model. A preliminary study (Antosiewicz et al., 1996) suggests that one of these parameter sets, PARSE (Sitkoff et al., 1994), improves the accuracy of pK_as calculated with the PB model when a low protein dielectric constant is used. The present paper presents a more extensive and systematic study of the degree to which the PARSE parameters improve accuracy. Of particular interest is whether these parameters make calculations with a protein dielectric constant of 4 as accurate as calculations with a protein dielectric constant of 20.

3. Conformation of the Protein. To date, most calculations of the pK_as of ionizable groups in proteins have used crystallographically determined structures. The use of crystal structures may be a source of error, however, because the average conformation of the protein in the crystal may differ from that in solution. Furthermore, a single conformation may be an inadequate representation of the space of conformations sampled by a protein at room temperature. One way to address this problem is to generate conformations that might be more "solution-like" than the starting crystal conformation, by computational methods. Accordingly, two studies have considered alternate conformations generated by MD simulations (Bashford & Gerwert, 1992; Yang & Honig, 1993). However, these methods have not been shown to yield improved accuracy. Moreover, it has been pointed out that the conformations generated by the straightforward application of MD will tend to be ones that stabilize the ionization states selected for use in the simulation (Bashford

& Gerwert, 1992; Yang & Honig, 1993). Therefore, pK_as computed with these conformations may be biased by the ionization states arbitrarily chosen for the simulation. A recent publication decribes the computation of pK_as, with sytematic variation of the conformations of ionizable side-chains around their initial crystal positions (You & Bashford, 1995). This procedure increases the accuracy of pK_as for HEWL. On the other hand, the accuracy still appears to be less than that of the trivial null model, which assumes that the protein does not shift pK_as at all (Antosiewicz et al., 1994). Also, this approach is still based upon a crystal structure, rather than a solution structure. It would appear that assessment of the influence of crystallization and conformational flexibility upon computed pK_as requires unbiased sets of protein conformations appropriate to the solution state.

The ensembles of conformations that result from NMR studies of proteins in solution may be helpful in this regard. Although they are frequently generated by simulated annealing procedures based upon MD, the conformations generated are constrained by the experimental data. They therefore represent the best available information on the solution structures of proteins. The present study therefore compares the accuracy of pK_as computed as averages over sets of conformations determined by NMR with the accuracy of pK_as computed with crystal structures.

B. Ion Binding in Ribonuclease A and pK_a Shifts in Ovomucoid Third Domain

The chief purpose of the present study is to address the questions raised in the preceding subsections, through comparisons among a large set of pK_a calculations on four proteins for which pK_as have been measured and for which solution conformations have been determined by NMR spectroscopy. However, some of the individual calculations are interesting in their own right. In particular, several pK_as in RNase A have been measured as a function of the concentration of phosphate. Because RNase A binds phosphate at a known site, it is of interest to determine whether the PB model is able to reproduce the measured pK_a shifts that occur upon binding. This is a case of the more general problem of predicting protonation changes associated with the noncovalent binding of small molecules by proteins. Also, Robertson and co-workers (Schaller & Robertson, 1995; Swint-Kruse & Robertson, 1995), propose explanations of the pK_a shifts they observe for carboxylic acid groups in OMTKY3. However, the explanations provided by the present calculations differ significantly.

C. Statistical Analysis of Measured pK_as

The number of carefully measured pK_as for ionizable groups in proteins is large enough that a statistical analysis is feasible. Here, the average and variance of the measured pK_as for each type of group (e.g., Glu, His) are computed. This analysis has two consequences. First, it is found that Asp side chains are, on average, considerably more acidic than Glu side chains. Second, the analysis of measured pK_as yields a new "null" model for the prediction of protein pK_as: this is the assumption that the pK_a of each group equals the mean measured pK_a for groups of the same type. This new null model is more accurate than the original null model which assumes that proteins do not shift pK_as at all. Nonetheless, the more accurate computational models described here are more predictive than the new null model.

METHODS

A. Theory and Computation

The computational methods used here have been described previously. Briefly, the pK_a of an ionizable group is defined as the pH for which it is half-ionized. The charge states are computed with either of two algorithms (Antosiewicz & Porschke, 1989; Gilson, 1993) that are based upon statistical thermodynamic theory described elsewhere (Schellman, 1975; Bashford & Karplus, 1990; Gilson, 1993). This theory requires as input the interaction energies among the ionizable groups of the protein, and the difference between the energy of ionizing each group in the otherwise neutral protein, relative to the energy of ionizing it in bulk solvent. These energies are all assumed to be electrostatic in nature, and are computed by the PB model (Warwicker & Watson, 1982; Gilson et al., 1985, 1988; Klapper et al., 1986; Gilson & Honig, 1988a; Honig et al., 1993; Madura et al., 1994). These energy calculations, in turn, take as inputs the structure of the protein and a number of other parameters. These inputs are discussed next.

When a group ionizes, its net charge and its charge distribution both change. Two approaches to the treatment of the charge changes that occur upon ionization are examined here. The simpler approach uses the full set of partial atomic charges appropriate to the neutral state of each group as a starting point. Ionization is modeled as addition or subtraction of a unit charge to one atom of the group. The atomic charges and radii used with this "single-site" model are the same as those used in a previous study (Antosiewicz et al., 1994): the charges are based upon the CHARMm Version 22.0 polar-hydrogen-only parameters (Molecular Simulations Inc., Waltham, MA; CHARMm Version 22.0, 1992), and the radii are based upon the OPLS nonbonded parameter set (Jorgensen & Tirado-Rives, 1988). The more detailed approach involves replacing the partial atomic charges appropriate to the neutral state with a completely different set of charges appropriate to the ionization state (Bashford & Gerwert, 1992; Yang et al., 1993). The present implementation of this "full-charge" model is detailed elsewhere (Antosiewicz et al., 1996). Here, the full-charge model is used with the PARSE atomic parameter set (Sitkoff et al., 1994).

Before the electrostatics calculations can be carried out, coordinates must be established for hydrogens. X-ray crystallography of proteins does not yield these coordinates, so they must be added by some computational procedure. Although hydrogen coordinates typically are included with NMR structure sets, the numbers and positions of hydrogens on ionizable groups are not typically determined by the data, so these coordinates are rather arbitrary. Using these coordinates as provided would make it difficult to compare results from crystal structures with results from NMR structures. Therefore, hydrogen coordinates are stripped from the NMR data sets, and recalculated by the same procedure used for the crystallographic structures. The following procedure is used. First, tautomeric forms for neutral histidines and rotational orientations for carboxylic acids are chosen. A neutral histidine can be protonated on

Table 1: Measured pK_a Values of Histidines in RNase A

solvent (nominal)	reference	histidine 12	48	105	119
	In the Absence of Phosphate[a]				
<100 mM NaCl	Rüterjans & Witzel, 1969	5.1	na	6.4	5.6
100 mM NaCl	Cohen et al., 1973	6.1	na	6.8	6.3
200 mM NaCl	Rüterjans & Witzel, 1969	5.8	na	6.5	6.1
200 mM NaCl	Matthews & Westmorland, 1973	6.0	na	6.6	6.2
200 mM sodium acetate	Walters & Allerhand, 1980	5.7	6.2	6.6	6.0
200 mM sodium acetate	Meadows et al., 1968	5.8	6.4	6.7	6.2
300 mM NaCl	Markley, 1975	5.8	6.3	6.7	6.2
400 mM NaCl	Rüterjans & Witzel, 1969	6.1	na	6.6	6.3
	In the Presence of Phosphate[a]				
100 mM NaCl, 6.5 mM PO_4	Cohen et al., 1973	6.4	na	6.8	6.6
100 mM NaCl, 20 mM PO_4	Cohen et al., 1973	6.8	na	6.8	7.0
100 mM NaCl, 65 mM PO_4	Cohen et al., 1973	6.8	na	6.8	7.1
100 mM NaCl, 100 mM PO_4	Cohen et al., 1973	7.2	na	7.0	7.6
200 mM NaCl, 80 mM PO_4	Meadows et al., 1969	6.6	na	6.7	6.9
200 mM NaCl, PO_4[b]	Rico et al., 1991	6.2	6.0	6.7	6.1

[a] NMR peak assignments are those of Markley (1975). [b] Measurement carried out in the presence of phosphate ion (Jorge Santoro, personal communication) at unspecified concentration.

either ND1 or NE2. The default here is to assume the proton is on ND1, but alternatives are considered in many cases, as noted in the Results. A neutral carboxylic acid can bear a proton on either of the oxygens for which coordinates are provided in the structure file. Here, the default is to place the proton on the oxygen that occurs second in the file of atomic coordinates. For the C-termini of peptides, the proton is linked to the OXT atom in the neutral form. Deviations from this practice are noted in the text. Once these choices have been made, the HBUILD (Brunger & Karplus, 1988) command of CHARMm Version 22.2 (Brooks et al., 1983) is used to establish energetically reasonable hydrogen positions. (Note that the HBUILD algorithm in Version 23.1 yields hydrogen coordinates that are noticeably different from those generated by Version 22.2.) Finally, except as otherwise noted, the energy of the system is minimized with respect to the positions of the hydrogens only by 500 steps of steepest-descent energy-minimization with CHARMm.

B. Atomic Coordinates

For HEWL, the triclinic crystal structure with Protein Data Bank (Bernstein et al., 1977) accession code 2LZT (Ramanadham et al., 1981) is used, and a set of 16 NMR structures was kindly provided by Drs. Lorna Smith and Christopher Dobson. For RNase A, the crystal structure complexed with sulfate ion, 3RN3 (Howlin et al., 1989), is used, and the 32 NMR structures are those listed under accession code 2AAS (Santoro et al., 1993). In calculations of the influence of phosphate upon the pK_as of ribonuclease, the phosphate group is placed at the location of the sulfate ion in crystal structure 3RN3. This is reasonable because the sulfate is virtually superimposable on the phosphate seen in the lower-resolution crystal structure 5RSA (Wlodawer et al., 1986). Crystal structure 3RN3 provides two sets of coordinates for His 119; calculations are done for both conformations, as noted in the text. Of the NMR structures, half have His 119 in one conformation, and half have His 119 in the other. For BPTI, the crystal structure used is 4PTI (Marquart et al., 1983), and the 20 NMR structures are from file 1PIT (Berndt et al., 1992). For OMTKY3, the crystal structure is from file 1PPF (Bode et al., 1986). This structure is not optimal, because OMTKY is complexed with human leukocyte elastase; however, no other complete crystal structure of OMTKY3 appears to be available. The 12 NMR structures of OMTKY are in file 1TUR.

C. Experimental pK_a Data and Solvent Conditions

For HEWL, the measured pK_as are those used previously (Antosiewicz et al., 1994), except where more recent NMR measurements are available (Bartik et al., 1994). These recent measurements were carried out at a uniform ionic strength of 100 mM, and this ionic strength was used for all calculations on HEWL. Some of the other measurements were made at higher ionic strengths, but it has been found that computed pK_as depend only weakly upon ionic strength above 100 mM (Antosiewicz et al., 1994).

For RNase A, most of the data were measured in a single NMR study in 200 mM NaCl (Rico et al., 1991). Accordingly, all calculations for RNase A assume an ionic strength of 200 mM. A number of additional pK_a measurements are available for the four histidine side chains (Table 1). The measured pK_as for the four histidines are taken to be simple averages of their pK_as measured at an ionic strength of 200 mM in the absence of phosphate. The resulting values for His 12, His 48, His 105, and His 119 are 5.8, 6.3, 6.6, and 6.1, respectively. The uncertainty in these values is $\sim\pm0.1$. [Note that although the full titration of His 48 is observed in the presence of acetate ion, this is not the case in simple NaCl (Markley, 1975). This issue is not addressed in the present work.] The addition of phosphate to the buffer makes His 12 and His 119 more basic (see Table 1). That His 12 and His 119 are strongly affected is consistent with crystallographic studies showing phosphate and other anions bound to the side chains of these two groups (Borah et al., 1985; Wlodawer et al., 1986; Howlin et al., 1989).

The Results section analyzing the influence of bound phosphate upon the pK_as in RNase A makes use of measurements in the presence of phosphate (see Table 1). Unfortunately, the available experimental data may not be completely adequate. This is because the computations assume that the phosphate binding site is fully occupied, but it is not certain that experimental data exist for saturating concentrations of phosphate: the pK_as of His 12 and His 119 rise between phosphate concentrations of 65 and 100

mM (Cohen et al., 1973) (see Table 1), and there do not appear to be any measurements for phosphate concentrations greater than 100 mM. As a consequence, the experimental pK_a shifts of these two residues upon the binding of phosphate probably should be viewed as lower bounds. The calculations for bound phosphate are compared with the highest experimental values. Note that although the pK_a measurements of (Rico et al., 1991) were carried out in the presence of phosphate ion (Jorge Santoro, personal communication), the pK_as of His 12 and His 119 do not appear to be shifted upward as in the other measurements listed in Table 1.

Here, calculations which include a phosphate ion treat the phosphate as an additional titratable group having an initial pK_a of 6.2. This corresponds to the second ionization of phosphoric acid. The net charge of "un-ionized" phosphate is therefore set to -1. The neglect of the ionizations at higher and lower pH is justified by the fact that we are chiefly interested in the influence of phosphate upon the pK_as of His 12 and His 119, which titrate near neutral pH. (When a protein dielectric constant of 20 is used, the computed pK_as of the second ionization of the phosphate range between 3.3 and 3.6, depending upon the histidine tautomer model used. This corresponds to a pK_a shift of about $3.5 - 6.2 = -2.7$. Application of this shift to the third ionization of the phosphate would not bring its initial pK_a of 12.7 into the relevant range.) The phosphorus atom and oxygen atoms are assigned radii of 1.87 Å and 1.48 Å respectively. These radii are based upon the OPLS nonbonded parameter set (Jorgensen & Tirado-Rives, 1988). The oxygens are treated as uncharged, and the phosphorus carries either 1e or 2e, depending upon its ionization state.

For BPTI, the measured pK_as are from NMR studies (Brown et al., 1976, 1978; Richarz & Wüthrich, 1978). The issue of ionic strength in these measurements is discussed elsewhere (Antosiewicz et al., 1994). Here, an ionic strength of 150 mM is used.

For OMTKY3, the measured pK_as of six acidic groups are from NMR studies (Schaller & Robertson, 1995; Swint-Kruse & Robertson, 1995). The measurements were carried out at nominal ionic strengths of 10 mM and 1 M. However, the buffers with a nominal ionic strength of 10 mM probably had ionic strengths of up to 50 mM (Schaller & Robertson, 1995). Except as otherwise noted in the text, the calculations are carried out at 10 mM ionic strength and 25 °C and compared with the corresponding experimental data. Sample calculations at 100 mM lead to no significant change in the qualitative conclusions reached in the present study. For some of the groups in OMTKY3, the experimental paper lists different apparent pK_as for several different protons. The data used here correspond to the proton(s) closest to the atom that actually loses the proton. For Asp 7, the apparent pK_as for C_β H (2.9) and C_β H′ (2.5) are averaged to yield a value of 2.7. Also, the graph of the chemical shift of the C-terminus against pH shows no plateau at low pH (Schaller & Robertson, 1995), so the reported value of 2.7 appears to be an upper limit. Therefore, any computed value of <2.7 is taken to be correct in the present study. Although the titration curves for the aspartic acid side chains of OMTKY3 are less complete than those for the glutamic acids, they appear to be sufficient to justify use of the reported values. Finally, it is worth noting that the pK_a of His 52 has been measured for chicken ovomucoid third domain, but it is not clear that the pK_a for the corresponding group in the OMTKY3 is the same.

The thickness of the Stern layer (Gilson & Honig, 1988b) is set to 2.0 Å in all calculations. Computed pK_as are insensitive to this parameter (Antosiewicz et al., 1994).

The use of deuterated water in some experiments is unlikely to affect the results by more that $0.1-0.2$ pK_a units, because of the consistent cancellation of isotope effects on the pH electrode and on the pK_as of ionizable groups (Glasoe & Long, 1960; Högfeldt & Bigeleisen, 1960; Bell & Kuhn, 1963).

RESULTS

A. Comparison of PARSE and Single-Site Parameters

The PARSE set of atomic charges and radii, unlike the single-site parameters that we have used heretofore, was developed and parameterized specifically for use with the PB electrostatics model. The parameterization is based upon the measured hydration energies of small molecules possessing the same chemical groups as found in proteins (Sitkoff et al., 1994). It is therefore reasonable to expect that the PARSE parameters will yield more accurate ionization energies when used in calculations of protein pK_as, and thus more accurate pK_as. This is one of the hypotheses addressed in this paper.

As a preliminary, it is confirmed that electrostatic energies computed in this laboratory with PARSE and the program UHBD (Davis et al., 1991), agree with published values computed with the program DelPhi (Sitkoff et al., 1994; Nicholls & Honig, 1991). Table 2 presents electrostatic dehydration energies for seven ionized and 20 neutral compounds, computed with the PARSE parameters and a solute dielectric constant of 2. The reported electrostatic energies neglect the work of forming a nonpolar cavity, and therefore are not directly comparable with experimental data. Previously published results are compared with two sets of dehydration energies computed here with the same atomic coordinates (Sitkoff et al., 1994). These two sets differ in the level of detail used to describe the molecular surface (Gilson et al., 1988). Energies computed with UHBD and the more detailed surface representation agree well with the published values. Energies computed with the less detailed surface representation deviate somewhat more. This less detailed representation is used for the pK_a calculations, because it requires less computer time. However, in the actual pK_a calculations, the discrepancies will be smaller than shown in Table 2, because the transfers are not to vacuum ($\epsilon = 1$) but to media of dielectric constant 4 or 20, and the magnitudes of any discrepancies are correspondingly reduced. (See Table 3, for example.)

In Table 3, the PARSE and single-site parameter sets are compared according to their predictions for the increase in the electrostatic energy of ionization when a side chain is transferred from water to the protein interior, with its lower dielectric constant. The changes in ionization energy for the single-site model are significantly greater than those for the PARSE parameters, except for the carboxylic acid group of Asp. The differences are especially marked for transfer to a dielectric constant of 4. It is shown below that these differences lead to significant, systematic errors in computed pK_as.

Table 2: Electrostatic Dehydration Energies of Small Molecules[a]

molecule	PUB	UHBD I	UHBD II
acetate	345.	343.	340.
propionate	340.	337.	334.
p-cresol ion	323.	320.	317.
methythiolate	328.	328.	326.
N-butylammonium	300.	299.	298.
N-*p*-guanidinium	287.	283.	281.
methylimidazolium	279.	276.	275.
methanol	30.1	29.3	28.9
ethanol	29.3	28.5	28.0
butanol	28.9	28.5	27.6
2-butanol	27.2	25.5	24.7
3-methyl-1-butanol	29.3	28.5	27.6
acetamide	49.4	47.3	46.4
propionamide	48.5	46.4	45.6
acetic acid	36.0	34.7	33.9
propionic acid	36.0	34.3	33.5
butyric acid	35.6	33.9	33.1
N-butylamine	27.6	26.8	26.4
methylethylsulfide	15.5	15.1	15.1
methylimidazole	51.9	50.6	49.8
methylthiol	13.8	13.4	13.0
methylindole	35.1	33.9	33.1
N-propylguanidine	56.1	53.1	51.9
toluene	13.0	13.0	12.1
p-cresol	35.6	34.7	33.5
phenol	36.0	34.7	33.9
2-methylphenol	35.1	33.5	32.2

[a] Comparison of electrostatic dehydration energies (kJ/mol) for small molecules computed with the PARSE parameter set. PUB, as published (Sitkoff et al., 1994); UHBD I, computed with the program UHBD with a highly detailed representation of the molecular surface [2500 initial sphere points per atom (Gilson et al., 1988)]; UHBD II, same as UHBD I but with less detailed surface (500 initial points per atom). The solute dielectric constant is set to 2, and the transfers are from water (dielectric constant 80) to vacuum (dielectric constant 1). The UHBD calculations assume an ionic strength of 0. Boundary conditions are assigned with Coulomb's law and the dielectric constant of the solvent. Two focusing steps are used (Gilson et al., 1988), and the final grid spacing is 0.2 Å.

Table 3: Computed Electrostatic Ionization Energies of Side Chains[a]

		energy (kcal/mol)		
side chain	ϵ	PARSE	single	PARSE′
Asp	4	74.8	75.1	74.1
Asp	20	11.9	11.7	11.7
Lys	4	66.9	86.7	67.6
Lys	20	10.4	13.5	10.7
Cys	4	76.8	91.9	76.1
Cys	20	12.2	14.3	12.0
Tyr	4	73.2	104.0	71.1
Tyr	20	11.9	15.9	11.1
His	4	51.4	64.5	52.7
His	20	7.9	10.1	8.4
Arg	4	55.3	75.7	55.0
Arg	20	8.8	11.9	8.7

[a] Computed changes in the electrostatic component of ionization energy (kJ/mol) of side chains upon transfer from water (dielectric constant 80) to a medium having a dielectric constant, ϵ, of either 4 or 20. PARSE, PARSE parameters, solute dielectric constant of 2; single, single-site model, solute dielectric constant set to ϵ (see text); PARSE′, PARSE parameters, solute dielectric constant set to ϵ.

As noted above, PARSE was parameterized for a solute dielectric constant of 2. However, for technical reasons, the pK_a calculations require that the dielectric constant of each group equal that of the protein as a whole. For example, when the protein is assigned a value of 4, the dielectric constant of each ionizable group is also set to 4. Table 3 shows that this alteration to the PARSE model has little effect upon the computed electrostatic energies.

Table 4: PARSE Parameters and Accuracy with a Protein Dielectric Constant of 4[a]

	no. of	RMS			max		
protein	sites	PARSE	single-site	null	PARSE	single-site	null
HEWL	20	2.1	3.0	1.3	7.4	10.2	3.1
RNASE	16	2.6/2.3	4.2/4.0	0.86	6.5/6.6	13.7/13.7	2.0
BPTI	10	0.74	0.93	0.60	1.3	2.0	1.0
OMTKY3	6	2.1	1.6	1.1	4.3	3.6	1.7
overall	52	2.1/2.0	3.1/3.0	1.0	7.4	13.7	3.1

[a] Comparison of accuracy of pK_as computed with PARSE parameters and single-site parameters for four proteins. HEWL, Hen egg white lysozyme; RNASE, Bovine pancreatic RNase A; BPTI, bovine pancreatic trypsin inhibitor; OMTKY3, turkey ovomucoid third domain; no. of sites, number of ionizable groups for which experimental data are available; null, results of null model, in which pK_as are assigned their unshifted initial values [see Methods and Antosiewicz et al. (1994)]; RMS, root mean square error relative to measured pK_as; max, maximum absolute error relative to measured pK_as; overall, cumulative results for all four proteins. The two results for RNase A are for the two crystal conformations of His 119.

In summary, the present implementation of the PARSE parameters yields excellent agreement with published results; and, in comparison with PARSE, the single-site model substantially overestimates the increased energy cost of ionization associated with transferring most ionizable side chains from water to a low-dielectric medium.

B. Cumulative Analysis of pK_a Calculations

This section evaluates the accuracy of the pK_as computed for four proteins, with several combinations of parameters and protein conformations. Each subsection addresses a specific question. These studies bear upon the development of accurate models, and suggest answers to some questions about the physical mechanisms that set the pK_as of ionizable groups in proteins. The first and second subsections describe results for protein dielectric constants of 4 and 20 respectively.

1. Protein Dielectric Constant of 4. a. The PARSE parameter set yields more accurate pK_as than the single-site parameters when used with a protein dielectric constant of 4 and crystal conformations. A recent study reports the pK_as computed with a protein dielectric constant of 20 are more accurate than those computed with a protein dielectric constant of 4 (Antosiewicz et al., 1994). This result is surprising, because the value of 4 is probably more realistic (Gilson & Honig, 1986; Simonson et al., 1991). One possible explanation is that the high protein dielectric constant served to compensate for inaccuracies in the single-site ionization model that was used. If this is true, then using a more realistic set of parameters, such as PARSE, might improve the accuracy.

The cumulative comparison in Table 4 shows that PARSE does indeed yield more accurate pK_as. OMTKY3 is the only protein for which PARSE decreases the accuracy, and the change is small. Still, as shown in the table, the PARSE results are still less accurate than the trivial null model of zero pK_a shifts.

At least part of the reason that PARSE parameters improve the accuracy is that the single-site parameters overestimate ionization energies, as reported above. The signature of this error is that the single-site calculations tend to predict that

Table 5: NMR Structures and Accuracy with a Protein Dielectric Constant of 4[a]

protein	no. of sites	RMS			max		
		NMR	cryst	null	NMR	cryst	null
HEWL	20	2.4	3.0	1.3	5.1	10.2	3.1
RNASE	16	4.1	4.2/4.0	0.86	11.	13.7/13.7	2.0
BPTI	10	0.65	0.93	0.60	1.0	2.0	1.0
OMTKY3	6	0.78	1.6	1.1	1.4	3.6	1.7
overall	52	2.7	3.1/3.0	1.0	11.	13.7	3.1

[a] Comparison of accuracy of pK_as based upon single crystal conformations, with those averaged over NMR structure sets. All calculations use single-site parameters and a protein dielectric constant of 4. See Table 4 for symbols. His 119 is in conformation A in one half of the NMR conformations, and in conformation B in the other half.

Table 6: PARSE/NMR Structures and Accuracy with a Protein Dielectric Constant of 4[a]

protein	no. of sites	RMS		max	
		NMR/PARSE	null	NMR/PARSE	null
HEWL	20	1.6	1.3	4.7	3.1
RNASE	16	5.6/6.0	0.86	21.	2.0
BPTI	10	0.71	0.60	1.5	1.0
OMTKY3	6	0.68	1.1	1.2	1.7
overall	52	3.3/3.5	1.0	21.	3.1
excl. RNASE	36	1.1	0.93	4.7	3.1

[a] Comparison of accuracy of pK_as computed with PARSE parameters, NMR structures, and a protein dielectric constant of 4, with accuracy of the null model. The two values for RNase A are for the two possible neutral tautomer states of histidines. See Table 4 for symbols; excl. RNASE, cumulative statistics excluding RNase A.

all groups are harder to ionize than they actually are: the pK_as of acids tend to be too high, and those of bases tend to be too low. This is apparent when one computes the following error gauge

$$S \equiv \sum_{i=1}^{N_{\text{group}}} z_i(\text{p}K_{ai}^{\text{calc}} - \text{p}K_{ai}^{\text{expt}}) \quad (1)$$

where z_i is -1 for acids and $+1$ for bases; N_{group} is the number of groups examined (52 here); and $\text{p}K_{ai}^{\text{calc}}$ and $\text{p}K_{ai}^{\text{expt}}$ are the calculated and measured pK_a values, respectively. If the errors in computed pK_as were random in sign, then S would be expected to have a value of 0. However, for the single-site model, $S = 1.2$ pK_a units. For the PARSE model with a protein dielectric constant of 4, $S = 0.13$ pK_a units.

Changing the assumed neutral tautomer of the histidines in these proteins does not change the conclusions presented in this section (results not shown).

b. Averaging over NMR conformations improves the accuracy of pK_as computed with a protein dielectric constant of 4 and single-site parameters. Another possible explanation for the observation that a protein dielectric constant of 20 yields more accurate pK_as than a protein dielectric constant of 4 is that this adjustment compensates for the use of a single crystal conformation of the protein, rather than an ensemble of solution conformations. Solution conformations are clearly more appropriate, because pK_as are measured for proteins in solution, rather than in a crystal. The accuracy comparison in Table 5 shows that pK_as computed as averages over the NMR structures yield somewhat more accurate pK_as than do the crystal structures. However, the cumulative improvement is not dramatic, and the NMR results still are not as accurate as the null model. The overall improvement might be more clear-cut if larger numbers of NMR conformations were used for each protein, because this would reduce the noise in the averaged pK_as: the RMSDs from the mean are 2.3, 1.6, 1.5, and 1.0, for HEWL, RNase A, BPTI, and OMTKY3, respectively. The NMR results are most accurate for BPTI and OMTKY3, the proteins for which the NMR pK_as have the least variation.

The improvement on going from crystal to NMR structures is most marked for OMTKY3: the RMSD drops from 1.6 to 0.78 pK_a units, and the maximum error drops from 3.6 to 1.4. This result is not entirely unexpected, because OMTKY3 is the only protein whose crystal structure was determined with the protein bound to another protein. Although graphical comparison of the structures does not reveal dramatic conformational differences, the binding of this protein to elastase may cause conformational readjustments greater than those that would result from crystallization alone.

The error gauge S (see above) is not reduced by use of NMR calculations; its values are 1.2 and 1.7, for single-site parameters with crystal structures and NMR structures, respectively. When the highly inaccurate pK_as of RNase A are excluded (see Table 5), the same qualitative conclusion is reached: the values of S for crystal and NMR structures are 0.81 and 0.80, respectively. It does not appear that the increased accuracy of the NMR results is a consequence of removal of any systematic error. It is tempting to argue that the NMR structures yield more accurate results because they are solution structures, and pK_as are measured in solution.

c. Combining PARSE parameters and NMR structures produces mixed effects upon accuracy, for a protein dielectric constant of 4. It might be hoped that combined use of PARSE parameters and NMR conformations would yield accuracy comparable with that of the null model, at least. The cumulative results in Table 6 demonstrate that this is not the case. However, closer examination reveals that when the poor results for RNase A are excluded from the cumulative results, the accuracy indeed approaches that of the null model. RNase A appears to be a particularly challenging test case for calculations with a protein dielectric constant of 4, as discussed below.

2. Protein Dielectric Constant of 20. Rather accurate pK_as are obtained when a protein dielectric constant of 20 is used with the single-site model and crystal conformations (Antosiewicz et al., 1994). Perhaps surprisingly, replacing single-site parameters by PARSE parameters does not substantially alter accuracy when the high protein dielectric constant is used, as shown in Table 7. However, Table 8 demonstrates that use of NMR conformations does yield pK_as slightly more accurate than those based upon crystal conformations. Also, the variation in computed pK_as for the various NMR conformations is significantly less than for a protein dielectric constant of 4: the RMSDs from the mean range between 0.36 and 0.56 pK_a units for the four proteins examined. Finally, as shown in Table 9, combining PARSE parameters and solution conformations does not significantly improve accuracy, relative to the single-site model with crystal conformations and a protein dielectric constant of 20.

Table 7: PARSE Parameters and Accuracy with a Protein Dielectric Constant of 20[a]

protein	no. of sites	RMS		max	
		PARSE	single-site	PARSE	single-site
HEWL	20	0.79	0.76	2.6	1.9
RNASE	16	0.76/.66	0.93/0.77	1.8/1.6	2.0/1.7
BPTI	10	0.49	0.48	0.92	1.0
OMTKY3	6	0.86	0.69	1.5	1.3
overall	52	0.74/0.71	0.77/0.71	2.6	2.0/1.9

[a] Comparison of accuracy of pK_as computed with PARSE parameters and the single-site model, for crystal conformations and a protein dielectric constant of 20. See Table 4 for symbols.

Table 8: NMR Structures and Accuracy with a Protein Dielectric Constant of 20[a]

protein	no. of sites	RMS		max	
		NMR	cryst	NMR	cryst
HEWL	20	0.88	0.76	2.0	1.9
RNASE	16	0.56	0.93/0.77	1.5	2.0/1.7
BPTI	10	0.33	0.48	0.65	1.0
OMTKY3	6	0.58	0.69	0.83	1.3
overall	52	0.67	0.77/0.71	2.0	2.0/1.9

[a] Comparison of accuracy of pK_as computed with NMR structures and crystal conformations, for the single-site model and a protein dielectric constant of 20. See Table 4 for symbols.

Table 9: PARSE/NMR and Accuracy with a Protein Dielectric Constant of 20[a]

protein	no. of sites	RMS		max	
		NMR/ PARSE	cryst/ single	NMR/ PARSE	cryst/ single
HEWL	20	0.94	0.76	2.5	1.9
RNASE	16	0.56	0.93/0.77	1.1	2.0/1.7
BPTI	10	0.41	0.48	0.83	1.0
OMTKY3	6	0.57/0.62	0.69	0.91/1.0	1.3
overall	52	0.71/0.72	0.77/0.71	2.5	2.0/1.9

[a] Comparison of accuracy of pK_as computed with PARSE parameters, NMR structures, and a protein dielectric constant of 20, with accuracy of the single-site model with crystal conformations. See Table 4 for symbols. Two values for RNase A are for the two conformations of His 119. Two values for OMTKY3 are for ionic strengths of 10 and 100 nM, respectively.

C. Analysis of Results by Protein

1. Hen Egg White Lysozyme. Because of its role in catalysis, Glu 35 is of particular interest. Its measured pK_a is shifted upward to 6.2. Although the single-site titration model with a protein dielectric constant of 20 is demonstrably accurate, this model does not correctly yield the upwardly shifted pK_a of Glu 35 (Antosiewicz et al., 1994). Published calculations with models that are less accurate overall have yielded essentially the correct shift (Bashford & Karplus, 1990; Yang et al., 1993). These calculations have assumed a protein dielectric constant of 4. Of the calculations presented here that use the crystal conformation, only that with the single-site model and a protein dielectric constant of 4 yield a realistic value for Glu 35, 5.8. However, all of the calculations with the NMR conformations and a protein dielectric constant of 4 yield mean pK_as that are shifted upward, to values from 7.3 to 9.8. This situation is not entirely satisfactory, because none of the calculations that use a protein dielectric constant of 4 displays the cumulative accuracy of the models based upon a protein dielectric constant of 20. However, it may be possible to achieve accurate results for Glu 35 as well as for the other groups in lysozyme: a recent study demonstrates that the computed pK_a of Glu 35 depends strongly upon which oxygen atom of the side chain is assumed protonated in the neutral form (Antosiewicz et al., 1996). Thus, it may well be that accuracy will increase when internal protonation equilibria of carboxylic acids, as well as of histidines, are included in the pK_a calculations. Further examination of the determinants of the pK_a of Glu 35 should account for experimental data on the roles of Trp 108 and Asp 52 (Inoue et al., 1992). These issues will be examined in future studies.

2. Ribonuclease A. a. Importance of tautomer equilibria. RNase A is the only protein of the four studied here for which the accuracy does not compare favorably with the null model when NMR structures are used along with the PARSE parameter set and a protein dielectric constant of 4 (see Table 6). The reason appears to be that RNase A possesses several ionizable groups that lie close together and are sequestered from solvent. This causes the ionization energies to be small sums of large desolvation and interaction energies of opposite sign. It also makes the results sensitive to the assumed neutral tautomers of the groups: when the proton in the neutral form of the histidines in RNase A is shifted from ND1 to NE2, with a protein dielectric constant of 4, the computed pK_a of His 48 changes from −0.2 to 10.8. In addition, the pK_a computed for Asp 14, which lies 4 Å from His 48, changes from 5.4 to 0.7. This means that changing the assumed neutral tautomer converts the His 48−Asp 14 interaction from a hydrogen bond to a salt-bridge. These results are for the crystal structure, PARSE parameters, and a protein dielectric constant of 4.

The complicated couplings among the ionizable groups make it difficult to predict the predominant neutral tautomers of the groups. It will therefore be desirable to establish a consistent, automated way of including tautomer equilibria in the titration calculations. This has been done already by Bashford et al. (1993) for histidines. It may prove just as critical to account for the internal protonation equilibria of neutral carboxylic acids, as noted in the section on HEWL. Perhaps accounting for these additional protonation states will improve the accuracy of the pK_as obtained with a low dielectric constant.

b. Binding of phosphate by RNase A. RNase A binds phosphate ion, and the experimental data summarized in Table 1 show that His 12 and His 119 become more basic as a consequence. It is of interest to determine how well the computational approach taken here reproduces the pK_a shifts due to phosphate. Initially, all histidines are assumed to be protonated at NE2 in their neutral form, but calculations are also presented for His 12 and His 119 assumed protonated at ND1 in their neutral states. The results for an assumed protein dielectric constant of 20 are presented in Table 10, and those for a protein dielectric constant of 4 in Table 11. (The computed pK_as in these tables differ slightly from those presented elsewhere in this paper because the hydrogen coordinates were not energy-minimized here.)

For an assumed protein dielectric constant of 20, including the phosphate causes His 12 to become more basic by 2.0−3.9 pK_a units, depending upon the assumed neutral tautomer. The measured pK_a shift upon adding 100 mM phosphate to the solvent is roughly 1.2 pK_a units. For His 119, the computed pK_a shifts are 0.7−2.6 pK_a units, and the measured

Table 10: Effect of Phosphate upon pK_As in RNase A, with Protein Dielectric Constant of 20[a]

	computed, cryst A				computed, cryst B				
	NE2-H		ND1-H		NE2-H		ND1-H		
group	no PO_4	PO_4	no PO_4	PO_4	no PO_4	PO_4	no PO_4	PO_4	expt
NTER	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.0	7.6
GLU 2	2.5	2.8	2.5	2.9	2.4	2.7	2.4	2.7	2.8
Glu 9	4.1	4.2	4.1	4.2	4.0	4.2	4.1	4.2	4.0
His 12	4.5	6.5	4.0	7.8	4.6	6.6	4.1	8.0	5.8/7.2
Asp 14	1.9	2.0	2.0	2.1	1.9	2.0	2.0	2.1	<2.0
Asp 38	2.8	2.9	2.8	2.9	2.7	2.9	2.7	2.9	3.1[a]
His 48	6.5	6.5	6.4	6.4	6.5	6.5	6.5	6.4	6.3
Glu 49	4.6	4.6	4.6	4.7	4.6	4.7	4.7	4.7	4.7
Asp 53	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.6	3.9
Asp 83	2.0	2.3	2.1	2.4	2.0	2.3	2.1	2.4	3.5
Glu 86	3.8	3.9	3.8	3.9	3.8	3.9	3.9	3.9	4.1
His 105	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.2	6.6
Glu 111	3.8	4.1	3.9	4.1	3.7	3.9	3.6	3.8	3.5
His 119	5.9	8.5	7.0	8.2	5.7	8.2	6.1	6.8	6.1/7.6
Asp 121	1.5	2.0	1.1	1.7	2.1	2.8	2.1	2.9	3.1
CTER	2.3	2.3	2.3	2.4	2.3	2.3	2.3	2.3	2.4
RMSD	0.69	0.56	0.83	0.56	0.61	0.45	0.66	0.46	
max abs	1.6	1.2	2.0	1.4	1.5	1.2	1.7	1.1	

[a] Calculated pK_as in RNase A, compared with experimental values. Except for the histidines, all experimental values are from (Rico et al., 1991). The histidine data are averages of measured pK_as at a nominal ionic strength of 200 mM, in the absence of phosphate ion (Rüterjans & Witzel, 1969; Patel et al., 1975; Meadows et al., 1968) (See Table 1). Computations use a protein dielectric constant of 20. Cryst A: computed using crystal structure with His 119 in conformation A; cryst B, same but with His 119 in conformation B; NE2-H, ND1-H, assumed protonation site for neutral forms of His 12 and His 119; RMSD, max abs: root mean square deviation and maximum absolute deviation from experiment; expt, measured pK_as; for His 12 and His 119, pK_as are given for the absence/presence of phosphate. Errors for computed pK_as in the absence/presence of phosphate are based upon the corresponding pK_as for His 12 and His 119.

Table 11: Effect of Phosphate upon pK_as in RNase A, with Protein Dielectric Constant of 4[a]

	computed, cryst A				computed, cryst B				
	NE2-H		ND1-H		NE2-H		ND1-H		
group	no PO_4	PO_4	no PO_4	PO_4	no PO_4	PO_4	no PO_4	PO_4	expt
NTER	6.7	6.7	6.7	6.7	6.7	6.7	6.7	6.7	7.6
Glu 2	2.0	2.2	2.0	2.1	1.9	2.1	1.9	2.0	2.8
Glu 9	4.9	4.7	4.8	4.7	4.8	4.7	4.8	4.7	4.0
His 12	−0.43	8.2	−0.15	14.4	−0.12	8.6	−0.28	14.8	5.8/7.2
Asp 14	7.7	7.2	7.5	7.2	7.7	7.1	7.5	7.2	<2.0
Asp 38	2.8	3.0	2.9	3.0	2.8	3.0	2.8	3.0	3.1[a]
His 48	−4.9	−4.8	−4.6	−4.5	−4.7	−4.8	−4.5	−4.6	6.3
Glu 49	6.6	6.5	6.6	6.6	6.7	6.5	6.6	6.6	4.7
Asp 53	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.9
Asp 83	3.4	3.4	3.2	3.5	3.3	3.3	3.2	3.5	3.5
Glu 86	4.9	5.0	4.9	5.0	4.9	5.0	4.9	5.0	4.1
His 105	4.3	4.2	4.3	4.3	4.3	4.3	4.3	4.3	6.6
Glu 111	4.5	4.7	4.5	4.7	4.5	4.6	4.4	4.5	3.5
His 119	6.0	12.0	9.1	11.0	6.0	11.9	5.9	7.0	6.1/7.6
Asp 121	0.70	1.1	−0.65	0.04	1.9	2.7	1.9	2.6	3.1
CTER	1.9	1.9	1.9	2.0	1.9	1.9	1.9	2.0	2.4
RMSD	3.7	3.4	3.7	3.8	3.6	3.4	3.5	3.7	
max abs	11.2	11.1	10.9	10.8	11.0	11.1	10.8	10.9	

[a] Same as Table 10, but calculations used a protein dielectric constant of 4. See Table 10 for symbols.

shift is roughly 1.4 pK_a units. Thus, the computed shifts of His 12 and His 119 are in the correct range, but they depend upon the choice of tautomer. Although the shifts for His 12 seem to be consistently excessive, the computations assume full occupancy of the phosphate binding site, a condition which may not be satisfied with 100 mM phosphate (see Methods).

With an assumed protein dielectric constant of 4, phosphate binding is predicted to cause unrealistically large shifts of 8−15 pK_a units in the pK_a of His 12. For His 119, excessive shifts are predicted for the NE2-H tautomer, but the shifts of 1−2 pK_a units for the ND1-H tautomer agree well with experiment. More generally, including the phosphate ion in the calculations does not increase the overall accuracy of the pK_as computed with a protein dielectric constant of 4.

The model of phosphate used here is rather crude, because all the charge is placed on the phosphorus and the oxygens are treated as nonpolar. Further calculations with more detailed charge models have also been carried out (results not shown). These models permit the user to select the phosphate oxygens assumed to be protonated. Varying the protonation sites assumed for $H_2PO_4^-$ causes the pK_a of His 12 to vary over about 1 pK_a unit, when a protein dielectric

Table 12: Two Proposed Explanations of the pK_a Shifts in OMTKY3[a]

	Asp 7	Glu 10	Glu 19	Asp 27	CTER 56
S+R	HB, Lys 34	Lys13	HB, Arg21	HB	disulfide; His52
comp	Lys13, Lys34	Lys13, Lys34	HB, Lys13, Arg21, Lys34	HB?, Lys29, His 52	Lys29, His52, Lys5

[a] Dominant influences upon the pK_as of carboxylic acids in turkey ovomucoid third domain, based upon analysis of structure and chemical shifts, S+R (Schaller & Robertson, 1995) and upon the present calculations (comp); HB, hydrogen bonding. Glu 43 is omitted because its pK_a is only slightly shifted.

Table 13: Analysis of pK_a Shifts in OMTKY3, with Single-Site Parameters, Crystal Structure, and Protein Dielectric Constant of 20[a]

	Asp 7	Glu 10	Glu 19	Asp 27	Glu 43	CTER 56
NTER 1	−0.10	−0.06	−0.05	−0.10	−0.05	−0.12
Asp 7	**−0.02**	0.49	0.14	0.08	0.07	0.05
Glu 10	0.49	**0.03**	0.19	0.06	0.07	0.04
Tyr 11	0.33	0.52	0.43	0.10	0.13	0.06
Lys 13	−0.25	−0.72	−0.30	−0.06	−0.09	−0.04
Glu 19	0.14	0.19	**−0.72**	0.08	0.08	0.08
Tyr 20	0.06	0.06	0.18	0.14	0.09	0.28
Arg 21	−0.11	−0.11	−0.31	−0.07	−0.05	−0.13
Asp 27	0.08	0.06	0.08	**0.38**	0.29	0.20
Lys 29	−0.05	−0.04	−0.07	−0.55	−0.29	−0.15
Tyr 31	0.09	0.07	0.11	1.64	0.43	0.16
Lys 34	−0.74	−0.59	−0.30	−0.09	−0.08	−0.06
Glu 43	0.07	0.07	0.08	0.29	**0.11**	0.07
His 52	−0.07	−0.05	−0.08	−0.22	−0.07	−0.76
Lys 55	−0.05	−0.05	−0.12	−0.10	−0.06	−0.34
CTER 56	0.05	0.04	0.08	0.20	0.07	**0.18**
comp pK_a	2.9	3.4	2.6	3.6	4.4	2.5
expt pK_a	2.7	4.1	3.2	2.3	4.8	<2.7

[a] Analysis of causes of the pK_a shifts of ionizable groups of known pK_a in OMTKY3, for single-site titration model, protein dielectric constant of 20, crystal structure. Entries are in units of pK_a shift. Diagonal terms give net effect of desolvation and neutral polar groups. Last two lines give computed and measured pK_as.

Table 14: Analysis of pK_a Shifts in OMTKY3, with Single-Site Parameters, NMR Structures, and Protein Dielectric Constant of 20[a]

	Asp 7	Glu 10	Glu 19	Asp 27	Glu 43	CTER 56
NTER 1	−0.14	−0.08	−0.05	−0.06	−0.04	−0.08
Asp 7	**−0.01**	0.65	0.12	0.07	0.08	0.06
Glu 10	0.65	**−0.20**	0.14	0.06	0.10	0.05
Tyr 11	0.38	0.80	0.26	0.09	0.13	0.06
Lys 13	−0.20	−0.48	−0.22	−0.06	−0.11	−0.04
Glu 19	0.12	0.14	**−0.35**	0.06	0.06	0.07
Tyr 20	0.06	0.06	0.15	0.15	0.09	0.32
Arg 21	−0.14	−0.13	−0.42	−0.07	−0.05	−0.12
Asp 27	0.07	0.06	0.06	**−0.11**	0.22	0.19
Lys 29	−0.05	−0.05	−0.06	−0.44	−0.18	−0.22
Tyr 31	0.08	0.08	0.08	0.96	0.40	0.14
Lys 34	−0.38	−0.50	−0.35	−0.07	−0.08	−0.07
Glu 43	0.08	0.10	0.06	0.22	**0.24**	0.06
His 52	−0.08	−0.06	−0.08	−0.24	−0.07	−0.84
Lys 55	−0.05	−0.05	−0.11	−0.10	−0.05	−0.49
CTER 56	0.06	0.05	0.07	0.19	0.06	**0.34**
comp pK_a	3.5	3.3	2.9	3.1	4.5	2.3
expt pK_a	2.7	4.1	3.2	2.3	4.8	<2.7

[a] Same as previous table, but entries are averages over results for the 12 NMR conformations of ovomucoid.

Table 15: Analysis of pK_a Shifts in OMTKY3, with PARSE Parameters, NMR Structures, and Protein Dielectric Constant of 4

	Asp 7	Glu 10	Glu 19	Asp 27	Glu 43	CTER 56
NTER 1	−0.15	−0.08	−0.06	−0.05	−0.03	−0.08
Asp 7	**0.11**	0.73	0.14	0.05	0.06	0.04
Glu 10	0.73	**0.02**	0.20	0.05	0.08	0.04
Tyr 11	0.25	1.17	0.23	0.05	0.07	0.03
Lys 13	−0.19	−0.64	−0.26	−0.05	−0.09	−0.03
Glu 19	0.14	0.20	**−0.07**	0.04	0.05	0.06
Tyr 20	0.03	0.03	0.08	0.11	0.07	0.18
Arg 21	−0.15	−0.17	−0.62	−0.05	−0.04	−0.11
Asp 27	0.05	0.05	0.04	**−0.91**	0.27	0.16
Lys 29	−0.03	−0.04	−0.05	−0.66	−0.20	−0.20
Tyr 31	0.04	0.04	0.04	1.30	0.33	0.07
Lys 34	−0.45	−0.87	−0.44	−0.05	−0.06	−0.05
Glu 43	0.06	0.08	0.05	0.27	**0.32**	0.05
His 52	−0.07	−0.05	−0.07	−0.25	−0.05	−0.87
Lys 55	−0.05	−0.04	−0.10	−0.10	−0.05	−0.59
CTER 56	0.04	0.04	0.06	0.16	0.05	**0.79**
comp pK_a	3.8	2.8	3.0	1.9	4.7	2.9
expt pK_a	2.7	4.1	3.2	2.3	4.8	<2.7

[a] Same as previous table, but computations based upon PARSE parameters and a protein dielectric constant of 4.

constant of 20 is assumed. It will therefore be of interest to create a more complete treatment of tautomer equilibria for phosphate ions.

3. Bovine Pancreatic Trypsin Inhibitor. The pK_as of BPTI are the least shifted of all the proteins examined here. Evidently, the ionizable groups in BPTI are in environments that strongly resemble bulk solvent. BPTI also is the protein for which all the nontrivial models work best (see Tables 4−9). Still, none of the models that use a protein dielectric constant of 4 beats the null model for BPTI, while all of the models that use a dielectric constant of 20 do.

4. Turkey Ovomucoid Third Domain. The protonation equilibria of this protein have recently been the subject of two thoughtful studies (Schaller & Robertson, 1995; Swint-Kruse & Robertson, 1995) that are the source of the pK_a data used here. These studies have been interpreted as suggesting that hydrogen bonds between carboxylic side chains and neutral polar groups—chiefly backbone NH groups—play a large role in shifting downward the pK_as of several carboxylic acids. The first line of Table 12 summarizes the suggested causes of the pK_a shifts of the groups whose pK_as were measured. It is of interest to compare these interpretations of the data with those provided by the accurate computational methods described in the present paper.

Three sets of computations are used here: the single-site model with a protein dielectric constant of 20, for both crystal and NMR structures; and the PARSE charges with a protein dielectric constant of 4 and NMR structures. All three sets of computations use an ionic strength of 10 mM. The RMS errors of these computations are 0.69, 0.58, and 0.69 pK_a units, respectively. The maximum errors are 1.3, 0.83, and 1.2 pK_a units, respectively. For the three models used, Tables 13−15 break down the influences upon the pK_as of the carboxylic acids. Each non-diagonal entry in these tables is the computed interaction energy between two groups, in units of pK_a shift. Each diagonal term is the pK_a shift when all other groups are assumed neutral. These diagonal terms include the effect of partial desolvation of each group by the low-dielectric protein and the net effect of all neutral polar groups. Therefore, the influence of hydrogen bonds appears in the diagonal entries, and the influence of charge−

Table 16: Distances among Ionizable Groups in OMTKY3[a]

	Asp 7	Glu 10	Glu 19	Asp 27	Glu 43	CTER 56
NTER 1	18.9	25.1	25.5	18.9	26.9	17.3
Asp 7	0.0	6.7	15.7	19.0	18.6	23.5
Glu 10	6.7	0.0	13.7	22.0	18.9	26.7
Tyr 11	9.0	6.7	8.5	17.0	14.4	21.2
Lys 13	10.0	4.8	10.3	20.6	16.7	25.1
Glu 19	15.7	13.7	0.0	19.0	18.9	19.1
Tyr 20	22.2	23.2	12.0	14.5	19.9	9.1
Arg 21	17.4	18.3	8.6	19.1	23.3	14.9
Asp 27	19.0	22.0	19.0	0.0	10.7	11.0
Lys 29	23.5	25.1	20.4	6.1	9.9	13.3
Tyr 31	17.7	19.8	16.4	3.5	8.2	12.3
Lys 34	5.0	6.3	10.7	17.7	17.5	20.8
Glu 43	18.6	18.9	18.9	10.7	0.0	20.5
His 52	20.0	24.1	18.6	10.8	20.4	4.9
Lys 55	24.2	26.0	15.3	16.5	23.6	7.9
CTER 56	23.5	26.7	19.1	11.0	20.5	0.0

[a] Distances (Å) between carboxylic acids of known pK_a and other ionizable groups in crystal form of OMTKY3. Distances based upon nitrogen atoms of amines, NE2 of histidines, CZ of arginines, carboxylic carbons of carboxylic acids, and oxygen of tyrosine side chains.

charge interactions appears in the off-diagonal entries. The last two lines of each table present the computed and measured pK_a of each group. If one sums the pK_a shifts in each column and adds the result to the initial pK_a of each group, one does not obtain precisely the computed pK_a at the bottom of the column. The reason is that these sums implicitly assume each group to be fully ionized, but some groups will not, in fact, be ionized at the pH where a given group titrates.

For the most part, the three tables yield the same qualitative conclusions, although there is some inconsistency in the diagonal terms, particularly for Asp 27. It appears that most of the downward pK_a shifts of the carboxylic acids result from interactions with cationic side chains. For example, the shift of Asp 7 is attributable primarily to interactions with Lys 13 and Lys 34. The diagonal term is essentially zero for this group, suggesting that hydrogen bonds have little influence upon its pK_a. In contrast, the negative diagonal terms for Glu 19 suggest that its pK_a is reduced significantly by hydrogen bonds. Inconsistent results are obtained for Asp 27: its diagonal terms range from −0.91 to 0.38. The spread of computed pK_as for this group probably is connected with the fact that it is partially sequestered from solvent. Table 12 summarizes the primary influences upon the pK_a of each group, based upon the cumulative results in Tables 13−15. Comparison with the qualitative conclusions of the experimental study shows that the computations ascribe a larger role to cationic groups, while the role of hydrogen bonds is found to be more modest.

Interestingly, the calculations yield accurate pK_as for the C-terminus without any need to invoke an inductive effect of the nearby disulfide bond (Schaller & Robertson, 1995): the pK_a shift of the C-terminus is attributed to substantial interactions with Lys 29, Lys 55, and especially His 52. That the experimental analysis ascribed little influence to Lys 29 and Lys 55 presumably results from their distances from the C-terminus: 13 and 8 Å respectively, in the crystal structure (Table 16). This emphasizes the long range of electrostatic interactions, particularly under the low-salt conditions examined here. Table 16 provides the distances among the ionizable groups analyzed in the previous tables. Correlation of these data with the interactions energies in Tables 13−15 reveals more cases of important interactions over ranges on the order of 8−10 Å, such as those of Glu 19 with Arg 21 and Lys 34.

Table 17: Influence of Ionic Strength upon pK_as in OMTKY3[a]

	comp				
group	10 mM	1 M	change	expt	atom
Asp 7	2.92	3.50	0.58	0.57	C β H′
Glu 10	3.37	3.91	0.54	0.25	C γ H
Asp 27	3.60	4.03	0.43	0.43/0.35	C β H/C β H′
Glu 43	4.35	4.49	0.14	0.30	C γ H
CSSC 56	2.47	3.30	0.83	∼0.1	C α/C β H

[a] Computed and experimental effects upon pK_as of increasing ionic strength from 10 mM to 1 M. Computations based upon single-site titration model with protein dielectric constant of 20. Experimental results based upon chemical shifts for listed atoms, at 35 °C.

If these long-ranged interactions are, indeed, important determinants of the pK_as of the carboxylic acids of this protein, then the shifts should diminish significantly with increasing ionic strength. However, the short-ranged influence of hydrogen bonds is expected to be insensitive to ionic screening. The experimental studies provide measured pK_as for five of the carboxylic acids at 1 M ionic strength. Table 17 compares measured and calculated pK_a changes on going from 10 mM salt to 1 M salt. These calculations use the single-site ionization model and the crystal structure. The calculated pK_a changes agree remarkably well with the measured changes for four of the five groups. The exception is the C-terminus, whose measured pK_a change is 0.1 pK_a unit and whose calculated pK_a change is 0.8 pK_a unit. It is not clear whether this discrepancy results from inaccuracies in the calculations, the experiments, or both. However, it is worth noting that the measured pK_a of Cys 56 is one of the least precise, because no titration plateau is observed at low pH (Schaller & Robertson, 1995).

The attribution of pK_a shifts to hydrogen bonds in the experimental study was based upon analysis of the pH dependence of chemical shifts (Schaller & Robertson, 1995). However, the calculations are not inconsistent with the chemical shift data. For one thing, the observed spectroscopic couplings between carboxylic acids and main-chain NH groups do not prove the existence of strong energetic interactions. The experimental data support this distinction, for the backbone NH groups of Leu 48 and Lys 55 do not form hydrogen bonds with ionizable groups, yet their chemical shifts change significantly with pH. In particular, it appears that the chemical shift of Leu 48 NH is coupled to the titration of Glu 43—the only group titrating near pH 5—although the groups lie 11 Å apart. It is unlikely that the NH group of Leu 48 stabilizes the ionized form of Glu 43 at this range. This case implies that a coupling between an NH group and a carboxylic acid need not indicate a stabilizing interaction.

Also, the diagonal terms in Tables 13−15 include both stabilizing energetic contributions from hydrogen bonding and destabilizing contributions from desolvation. Thus, a diagonal term near zero implies only that these terms are closely balanced; i.e., any destabilization due to desolvation is fully compensated by hydrogen bonding. In such a case, the hydrogen-bonding group will not produce a net pK_a shift, but the charged form of the carboxylic acid may well produce a strong electrostatic field at a nearby NH group, generating a chemical shift.

Table 18: Statistical Analysis of Experimental pK_as and Comparison with Accurate pK_a Calculations[a]

		experiment			calculations		
group	*N*	mean	RMSD	max abs err	mean	RMSD	max abs err
Asp	19	2.7	0.87	1.8	2.7	0.79	1.7
Glu	13	4.0	0.87	2.2	3.6	0.80	1.9
Lys	11	10.1	1.2	3.7	10.4	0.73	2.0
His	10	6.9	1.1	2.2	6.2	0.98	1.7
Tyr	3	10.7	0.99	1.4	10.2	0.67	1.1
NTER	3	8.2	0.49	0.63	7.8	0.67	1.0
CTER	4	2.7	0.19	0.30	2.7	0.30	0.58
cumulative	63		0.94	3.7		0.78	2.0

[a] Statistical analysis of accuracy of single-site model with a protein dielectric constant of 20, and of new null model in which the pK_a of each type of group is estimated as the mean of the measured pK_as for that type. *N*, number of groups in data set; mean, simple average of either computed or measured pK_as; RMSD, max abs err: root mean square and maximum absolute errors. For calculations, these refer to differences between computed and measured pK_as. For experiment, these refer to the difference between each measured pK_a and the mean pK_a for groups of the same type.

D. Statistical Analysis of Measured pK$_a$s: Another Null Model

The trivial null model (Antosiewicz et al., 1994) used up to this point consists of the assumption that proteins do not shift the pK_as of ionizable groups from their initial values. The initial values used here are as follows: Lys, 10.4; Arg, 12.0; Glu, 4.4; Asp, 4.0; His, 6.3; Tyr, 9.6; Cys, 8.3; peptide *N*-terminus, 7.5; peptide C-terminus, 3.8. These pK_as are close to those of small molecules bearing chemical groups of the same type. However, a more accurate null model can be constructed from the pK_as actually observed for these groups in proteins. The RMS deviations of these pK_as from the average measured pK_as indicate the width of the distribution of actual pK_as. Averages and deviations of this type are presented in Table 18. In order to maximize the size of the samples, data are drawn not only from the four proteins examined in this study, but also from chymotrypsin (Bender et al., 1964; Fersht & Sperling, 1973; Fersht & Requena, 1971; Fersht, 1972), T4 lysozyme (Anderson et al., 1990), a mutant of staphylococcal nuclease (Stites et al., 1991), and ribonuclease T1 (Inagaki et al., 1981; Shirley et al., 1989). These data are all included in a previous paper on the calculations of pK_as (Antosiewicz et al., 1994); that paper also discusses the uncertainty in the experimental values. In cases where a measured pK_a is known only to be greater than or less than some value, the value itself is used as the experimental pK_a.

The overall RMSD of the mean pK_as from the actual pK_as is 0.94 pK_a unit, and the maximum deviation is 3.7. Table 18 and Figure 1 compare this "accuracy" with the performance of one of the more accurate computational models described here, single-site atomic parameters with a protein dielectric constant of 20. Comparison of all the computed pK_as with experiment yields an RMS deviation of 0.78 pK_a unit and a maximum absolute error of 2.0. Thus, the computed pK_as are more accurate than the new null model. [Most of the calculations are based upon uniform use of the neutral histidine tautomer with ND1 protonated. However, a few of the results are drawn from previous study (Antosiewicz et al., 1994) which did not include this tautomer choice in every case. In such cases, the present analysis uses the average of the pK_as computed for the histidine tautomers that were actually used.] As shown in Table 19, linear regression analyses also indicate that the calculated pK_as are more accurate than the new null model. Furthermore, somewhat better agreement with experiment would be expected if NMR conformations were used in place of the crystal structures.

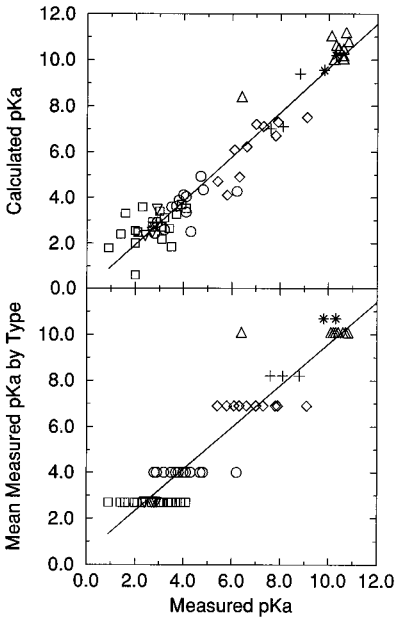

FIGURE 1: Scatter plots of calculated pK_as *vs* experiment and of mean pK_a by group *vs* experiment. Top: pK_as computed by single-site model with protein dielectric constant of 20. Bottom: pK_as estimated as mean of measured pK_a for each type of group. (△) Lys; (*) Tyr; (+) N-terminus; (◇) His; (○) Glu; (□) Asp; (▽) C-terminus.

Table 19: Linear Regression Analysis of New Null Model and Accurate pK_a Calculations[a]

	corr	slope	*Y*-intercept
new null model	0.96	0.91	0.52
computations	0.97	0.96	0.0

[a] Linear regression analysis of new null model and of computations with single-site model and a protein dielectric constant of 20. Columns show correlation coefficient, slope, and *Y*-intercept. Ideally, these would equal 1, 1, and 0. See also Figure 1.

The present analysis of experimental pK_as also shows that the mean pK_a of Asp side chains is shifted down to 2.7, well below its initial pK_a of 4.0. In contrast, the mean pK_a of Glu side chains, 4.0, is close to its normal value of 4.4. Figure 2 provides histograms of the distributions of the measured pK_as of Asp and Glu, and Table 20 shows that the tendency of Asp side chains to be more acidic than Glu side chains holds separately in each of the proteins examined here. That Asp pK_as tend to be lower than Glu pK_as could be explained in at least two ways. First, the intrinsic pK_a of an Asp side chain might be lower than is normally supposed. However, this seems unlikely, given that the pK_a of an Asp side chain in solution is about 3.9 (Stryer, 1981). Furthermore, the pK_as of acetic acid, propionic acid, and butyric acid are 4.75, 4.87, and 4.82, respectively (Martell & Smith, 1974). It appears that the protein environments of Asp residues—at least those examined here—tend to stabilize the ionized form quite significantly. If this is correct, then it would be reasonable to expect that the mean of the *computed* pK_as of Asp residues

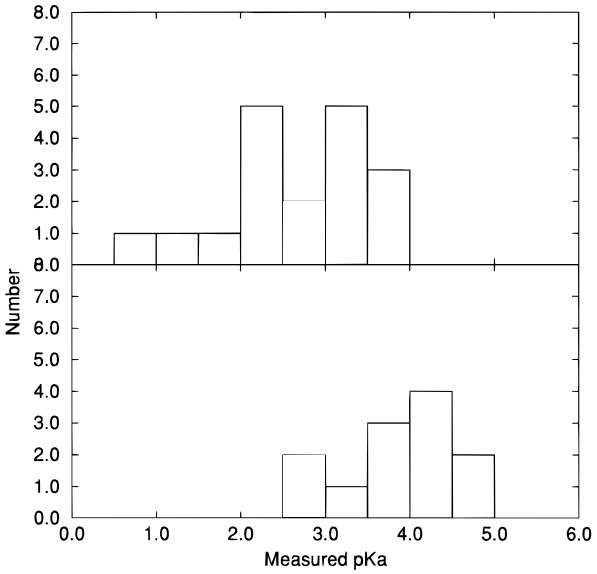

FIGURE 2: Distribution of measured pK_as for Asp (top) and Glu (bottom).

Table 20: Mean Experimental pK_as of Asp and Glu, According to Protein

	Asp		Glu	
protein	N^a	mean[b]	N	mean
HEWL	7	2.6	2	4.6
RNase A	5	3.1	5	3.8
BPTI	2	3.2	2	3.8
OMTKY3	2	2.5	3	4.0
CHYMO	2	2.0	0	na[c]
T4	1	2.0	0	na
RNase T1	0	na	1	4.3
cumul	19	2.7	13	4.0

[a] N, number of measured pK_as in each protein. [b] Mean: Average pK_a of group. [c] na, not applicable.

would be shifted down. This should not be observed for Glu residues, however. This is precisely what is observed: for Asp, the mean of the pK_as computed as described in the previous paragraph is 2.7; for Glu, the mean is 3.6. The computational model thus accounts extremely well for the observed differences in the environments of the two carboxylic acid side chains.

DISCUSSION

A. Calculating Accurate pK_as

Although using PARSE parameters or NMR structures improves the accuracy obtained when the protein dielectric constant is assumed to be 4, the most accurate results still are obtained for the high dielectric constant of 20. The results for a low protein dielectric constant may improve further with more detailed treatments of conformational flexibility and of tautomer equilibria. For now, however, accurate calculations require the use of a high dielectric constant. Use of PARSE parameters with a high protein dielectric constant produces little if any benefit, but use of NMR structures produces a small but fairly consistent improvement. However, it is not clear that the small benefit is worth the computational cost. Therefore, it is still reasonable to use single-site parameters and a protein dielectric constant of 20 in practical applications.

It is recently been pointed out by E. Mehler that the accuracy of this approach actually is slightly better than previously appreciated. The reason is that the overall RMSD previously reported by us (Antosiewicz et al., 1994) is based upon data for three different tautomer states of the same system, T4 lysozyme. This resulted in triple-counting of a particularly unfavorable case. When only one tautomer case is included, the overall RMSD drops from 0.89 to about 0.71.

B. Parameters, Conformations, and the Dielectric Constant of the Protein

The first part of the Results section indicates that the single-site atomic parameters used in our previous study overestimate the energetic cost of desolvating ionized groups in the protein interior. As a consequence, the single-site model also overestimates the energetic costs of ionization in the low dielectric protein, producing systematic errors. Use of the more realistic PARSE parameter set removes this systematic error, and improves the accuracy of the computed pK_as. Assigning the protein a high dielectric constant of 20 also makes this systematic error negligible. It therefore appears that one reason a protein dielectric constant of 20 yielded more accurate pK_as than a protein dielectric constant of 4 is that the high value removed a systematic error from the calculations. Unfortunately, even when the PARSE parameters are used, pK_as computed with a protein dielectric constant of 4 are less accurate than those computed with a less plausible protein dielectric constant of 20.

The present study shows that pK_a calculations based upon NMR structures tend to yield more accurate results than calculations based upon single crystal structures. The level of improvement varies from protein to protein, and upon the assumed protein dielectric constant. The greatest improvement is for OMTKY3, and the least is for RNase A. As suggested in Results, the RNase A is particularly complicated, because of the sensitivity of the results to tautomer states; and the crystal structure of OMTKY3 may be particularly inappropriate, because the protein of interest is complexed with another protein.

Still, the basis for the general improvement in accuracy upon using NMR structures is uncertain. One possibility is that the solution structures obtained by NMR yield better results because pK_as are measured on proteins in solution. On the other hand, as shown in the Appendix, the NMR structure sets yield rather large ranges of pK_as, especially when the dielectric constant of the protein is set to 4. Therefore, it is not clear that the solution structures are particularly well-defined with respect to pK_as. Perhaps, then, the accuracy of the NMR results has more to do with the sampling of a range of realistic conformations than with the use of a particularly accurate structure. This idea is consistent with the observation that limited conformational sampling around the crystal structure of HEWL improves the accuracy of computed pK_as (You & Bashford, 1995).

The use of NMR data might be improved in at least two ways. First, it should be possible to reduce the statistical uncertainty in the averaged pK_as by generating and averaging over more structures. Second, it should be possible to Boltzmann-weight the contributions of the various conformations (You & Bashford, 1995). In the present study, all the NMR conformations are accorded equal weights when average pK_as are computed.

C. Binding of Phosphate by RNase A

Phosphate binding by RNase A is a preliminary test of the ability of the models used here to predict the influence of ligand binding upon protonation states. It is gratifying that reasonable pK_a shifts are obtained when a protein dielectric constant of 20 is assumed. This is despite the fact that the ionizable groups whose pK_as shift are actually in contact with the phosphate and that the phosphate itself is titratable. More accurate results may well be obtained in less challenging systems. Also, it is not clear that the experimental data are adequate for the present study, because the phosphate binding site may not be fully occupied in the experiments.

The calculation of pK_a changes associated with ligand binding is of great practical importance and will be the subject of further study. For now, it seems likely that pK_a shifts computed for ionizable groups that are not actually in contact with a charged ligand will be rather accurate when a protein dielectric constant of 20 is used. This supports the validity of a previous study that examines the influence of cation binding upon the pK_a of the catalytic histidine in acetylcholinesterase (Wlodek et al., 1995).

D. pK_a Shifts in OMTKY3

The calculations for OMTKY3 are of interest because they yield a rather different explanation for the pK_a shifts of acidic groups than that obtained by graphical examination of the structure and interpretation of NMR data (Schaller & Robertson, 1995; Swint-Kruse & Robertson, 1995). The calculations place considerably more emphasis upon the remote influence of ionized groups and less upon short-ranged hydrogen bonding. The calculations suggest that replacement of Lys 13 by a neutral residue would make Asp 7, Glu 10, and Glu 19 noticeably less acidic and that neutralization of Lys 29 would have a similar but weaker effect upon Asp 27, Glu 43, and the C-terminus.

E. New Null Model

The present paper includes a new null model for the prediction of protein pK_as. In this model, the pK_a of a group is estimated as the mean of the observed pK_as for groups of the same type. For example, the pK_a of every Asp is estimated to be 2.7. The disadvantage of such a model is that it will vary with the data set used to compute the mean pK_as. On the other hand, it offers stringent competition for the computational method. As shown here, the single-site parameters with a dielectric constant of 20 and crystal structures outperform even this new null model.

F. The Protein Environment of Aspartic Acid

Unexpectedly, the survey of experimental pK_a data shows that aspartic acid is unique in possessing a substantially shifted average pK_a. The mean experimental pK_a of the 19 Asp residues in the present data set is shifted downward to 2.7. By contrast, the mean pK_a of the 13 Glu residues is 4.0. The mean pK_as of other types of ionizable group are, like Glu, unshifted. It is gratifying that the computations essentially reproduce these averages. It appears that Asp residues, but not Glu residues, tend to lie in environments that favor the ionized form. Although this might represent an idiosyncracy of the proteins examined here, it also could represent a general phenomenon and merits further investigation.

ACKNOWLEDGMENT

We are grateful to C. M. Dobson and L. J. Smith for providing the NMR structures of HEWL; to C. G. Hoogstraten, A. Krezel, and J. Markley for providing the NMR structures of OMTKY3 prior to publication; to M. Rico and J. Santoro for providing the NMR structures of RNase A prior to publication, and for additional information and helpful discussions; to A. Robertson for providing pK_a data and other information on OMTKY3 prior to publication; to B. Honig, K. Sharp, and D. Sitkoff, for providing molecular coordinates and parameter files for PARSE; to D. Gorman, E. Mehler, and N. Pace for helpful discussions; and to H. S. R. Gilson for thoughtful comments on the text. Certain commercial equipment or materials are identified in this paper in order to specify the methods adequately. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

SUPPORTING INFORMATION AVAILABLE

Lists of the experimental pK_as used for validation and most of the computed pK_as used in the analyses described in this paper (20 pages). Ordering information is given on any current masthead page.

REFERENCES

Anderson, D. E., Becktel, W. J., & Dahlquist, F. W. (1990) *Biochemistry 29*, 2403−2408.
Antosiewicz, J., & Porschke, D. (1989) *Biochemistry 28*, 10072−10078.
Antosiewicz, J., McCammon, J. A., & Gilson, M. K. (1994) *J. Mol. Biol. 238*, 415−436.
Antosiewicz, J., Briggs, J. M., Elcock, A. H., Gilson, M. K., & McCammon, J. A. (1996) *J. Comput. Chem.* (in press).
Bartik, K., Redfield, C., & Dobson, C. M. (1994) *Biophys. J. 66*, 1180−1184.
Bashford, D., & Karplus, M. (1990) *Biochemistry 9*, 327−335.
Bashford, D., & Gerwert, K. (1992) *J. Mol. Biol. 224*, 473−486.
Bashford, D., Case, D. A., Dalvit, C., Tennant, L., & Wright, P. E. (1993) *Biochemistry 32*, 8045−8056.
Bell, R. P., & Kuhn, A. T. (1963) *Trans. Faraday Soc. 59*, 1789−1793.
Bender, M. L., Clement, G. E., Kezdy, F. J., & d'A. Heck, H. (1964) *J. Am. Chem. Soc. 86*, 3680−3690.
Berndt, K. D., Guntert, P., Orbons, L. P. M., & Wüthrich, K. (1992) *J. Mol. Biol. 227*, 757.
Bernstein, F. C., Koetzle, T. F., Williams, T. F., Meyer, G. J. B., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535−542.
Beroza, P., & Fredkin, D. (1996) *J. Comput. Chem.* (in press).
Beroza, P., Fredkin, D. R., Okamura, M. Y., & Feher, G. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 5804−5808.
Beroza, P., Fredkin, D., Okamura, M., & Feher, G. (1995) *Biophys. J. 68*, 2233−2250.
Bode, W., Wei, A. Z., Huber, R., Meyer, E., Travis, J., & Neumann, S. (1986) *EMBO J. 5*, 2453.
Borah, B., Chen, C. W., Egan, W., Miller, M., Wlodawer, A., & Cohen, J. S. (1985) *Biochemistry 24*, 2058.
Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983) *J. Comput. Chem. 4*, 187−217.
Brown, L. R., Marco, A. D., Wagner, G., & Wüthrich, K. (1976) *Eur. J. Biochem. 62*, 103−107.

Brown, L. R., Marco, A. D., Richarz, R., Wagner, G., & Wüthrich, K. (1978) *Eur. J. Biochem. 88*, 87−95.
Brunger, A. T., & Karplus, M. (1988) *Proteins: Struct., Funct., Genet. 4*, 148−156.
Cohen, J. S., Griffin, J. H., & Schechter, A. N. (1973) *J. Biol. Chem. 248*, 4305−4310.
Davis, M. E., Madura, J. D., Luty, B. A., & McCammon, J. A. (1991) *Comput. Phys. Commun. 62*, 187−197.
Fersht, A. R. (1972) *J. Mol. Biol. 64*, 497−509.
Fersht, A. R., & Requena, Y. (1971) *J. Mol. Biol. 60*, 279−290.
Fersht, A. R., & Sperling, J. (1973) *J. Mol. Biol. 74*, 137−149.
Gilson, M. K. (1993) *Proteins: Struct., Funct., Genet. 15*, 266−282.
Gilson, M. K., & Honig, B. H. (1986) *Biopolymers 25*, 2097−2119.
Gilson, M. K., & Honig, B. H. (1987) *Nature 330*, 84−86.
Gilson, M. K., & Honig, B. (1988a) *Proteins: Struct., Funct., Genet. 4*, 7−18.
Gilson, M. K., & Honig, B. (1988b) *Proteins: Struct., Funct., Genet. 3*, 32−52.
Gilson, M. K., Rashin, A. A., Fine, R., & Honig, B. (1985) *J. Mol. Biol. 183*, 503−516.
Gilson, M. K., Sharp, K. A., & Honig, B. H. (1988) *J. Comput. Chem. 9*, 327−335.
Glasoe, P. K., & Long, F. A. (1960) *J. Phys. Chem. 64*, 188−190.
Hagler, A., Huler, E., & Lifson, S. (1974) *J. Am. Chem. Soc. 96*, 5319−5327.
Högfeldt, E., & Bigeleisen, J. (1960) *J. Am. Chem. Soc. 82*, 15−20.
Honig, B., Sharp, K., & Yang, A.-S. (1993) *J. Phys. Chem. 97*, 1101−1109.
Howlin, B., Moss, D. S., & Harris, G. W. (1989) *Acta Crystallogr., Sect. A 45*, 851.
Inagaki, F., Kawano, Y., Shimada, I., Takahashi, K., & Miyazawa, T. (1981) *J. Biochem. 89*, 1185−1195.
Inoue, M., Yamada, H., Yasukochi, T., Kuroki, R., Miki, T., Horiuchi, T., & Imoto, T. (1992) *Biochem. 31*, 5545−5553.
Jorgensen, W. L., & Tirado-Rives, J. (1988) *J. Am. Chem. Soc. 110*, 1657−1666.
Kendrew, J. (1963) *Science 139*, 1259−1266.
King, G., Lee, F. S., & Warshel, A. (1991) *J. Chem. Phys. 95*, 4366−4377.
Klapper, I., Hagstrom, R., Fine, R., Sharp, K., & Honig, B. (1986) *Proteins: Struct. Func. Gen. 1*, 47−79.
Krezel, A. M., Darba, P., Robertson, A., Fejzo, J., Macura, S., & Markley, J. (1994) *J. Mol. Biol. 242*, 203−214.
Langsetmo, K., Fuchs, J. A., Woodward, C., & Sharp, K. A. (1991) *Biochemistry 30*, 7609−7614.
Linderstrom-Lang, K. (1924) *C. R. Trav. Lab. Carlsberg 15*(7).
Madura, J. D., Davis, M. E., Gilson, M. K., Wade, R. C., Luty, B. A., & McCammon, J. A. (1994) *Rev. Comput. Chem. 5*, 229−267.
Mandel, M. (1964) *Proc. Natl. Acad. Sci. U.S.A. 52*, 736−741.
Mandel, M. (1965) *J. Biol. Chem. 240*, 1586−1592.
Markley, J. L. (1975) *Biochemistry 14*, 3546−3554.
Marquart, M., Walter, J., Deisenhofer, J., Bode, W., & Huber, R. (1983) *Acta Crystallogr., Sect. B 39*, 480.
Martell, A., & Smith, R. (1974) *Critical Stability Consants*, Vol. 1-IV, Plenum Press, New York.
Matthews, C. R., & Westmorland, D. G. (1973) *Ann. N.Y. Acad. Sci. 222*, 240−253.
Meadows, D. H., Jardetzky, O., Epand, R. M., Rüterjans, H. H., & Scheraga, H. A. (1968) *Proc. Natl. Acad. Sci. U.S.A. 60*, 766−772.
Meadows, D. H., Roberts, G. C. K., & Jardetzky, O. (1969) *J. Mol. Biol. 45*, 491−511.
Molecular Simulations Inc. Waltham, MA. (1992) CHARMm Version 22.0.
Nicholls, A., & Honig, B. (1991) *J. Comput. Chem. 12*, 435−445.
Oberoi, H., & Allewell, N. M. (1993) *Biophysical Journal 65*, 48−55.
Patel, D. J., Canuel, L. L., Woodward, C., & Bovey, F. A. (1975) *Biopolymers 14*, 959−974.
Ramanadham, M., Sieker, L. C., & Jensen, L. H. (1981) *Acta Crystallogr., Sect. A 37C*, 33.
Richarz, R., & Wüthrich, K. (1978) *Biochemistry 17*, 2263−2269.
Rico, M., Santoro, J., Gonzalez, C., Bruix, M., & Neira, J. L. (1991) Solution structure of bovine pancreatic ribonuclease A and ribonuclease-pyrimidine nucleotide complexes as determined by ^{1}H NMR, in *Structure, Mechanism and Function of Ribonucleases. Proceedings of the 2nd International Meeting held in Sant Feliu de Guíxols, Girona, Spain, 1990* (Cuchillo, C. M., de Llorens, R., Nogués, M. V., & Parés, X., Eds.) pp 9−14, Departament de Bioquímica i Biologia Molecular and Institut de Biologia Fonamental Vicent Villar Palasí, Universitat Autònoma de Barcelona, Bellaterra, Spain.
Rüterjans, H., & Witzel, H. (1969) *Eur. J. Bioch. 9*, 118−127.
Santoro, J., Gonzalez, C., Bruix, M., Neira, J. L., Nieto, J. L., Herranz, J., & Rico, M. (1993) *J. Mol. Biol. 229*, 722−734.
Schaller, W., & Robertson, A. D. (1995) *Biochemistry 34*, 4714−4723.
Schellman, J. A. (1975) *Biopolymers 14*, 999−1018.
Schmidt, A. B., & Fine, R. M. (1994) *Mol. Simul. 13*, 347−365.
Shirley, B. A., Stanssen, P., Steyaert, J., & Pace, C. N. (1989) *J. Biol. Chem. 264*, 11621−11625.
Simonson, T., & Perahia, D. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 1082−1086.
Simonson, T., Perahia, D., & Brunger, A. T. (1991) *Biophys. J. 59*, 670−690.
Sitkoff, D., Sharp, K. A., & Honig, B. (1994) *J. Phys. Chem. 98*, 1978−1988.
Smith, P. E., Brunne, R. M., Mark, A. E., & van Gunsteren, W. F. (1993) *J. Phys. Chem. 97*, 2009−2014.
Stites, W. E., Gittis, A. G., Lattman, E. E., & Shortle, D. (1991) *J. Mol. Biol. 221*, 7−14.
Stryer, L. (1981) *Biochemistry*, 2nd Ed., W. H. Freeman and Co., New York.
Swint-Kruse, L., & Robertson, A. D. (1995) *Biochemistry 34*, 4724−2732.
Takahashi, T., Nakamura, H., & Wada, A. (1992) *Biopolymers 32*, 897−909.
Walters, D. E., & Allerhand, A. (1980) *J. Biol. Chem. 255*, 6200−6204.
Warshel, A. (1981) *Biochemistry 20*, 3167−3177.
Warwicker, J., & Watson, H. C. (1982) *J. Mol. Biol. 157*, 671−679.
Wlodawer, A., Borkakoti, N., Moss, D. S., & Howlin, B. (1986) *Acta Crystallogr., Sect. B 42*, 379.
Wlodek, S. T., Antosiewicz, J., McCammon, J. A., & Gilson, M. K. (1995) Binding of cations and protons in the active site of acetylcholinesterase, in *Modeling of biomolecular structures and mechanisms* (Pullman, A., Jortner, J., & Pullman, B., Eds.) pp 25−37, Kluwer, Dordrecht, The Netherlands.
Yamazaki, T., Nicholson, L. K., Torchia, D. A., Wingfield, P., Stahl, S. J., Kaufman, J. D., Eyermann, C. J., Hodge, C. N., Lam, P. Y. S., Ru, Y., Jadhav, P. K., Hwan Chang, C., & Weber, P. C. (1994) *J. Am. Chem. Soc. 116*, 10791−10792.
Yang, A.-S., & Honig, B. (1993) *J. Mol. Biol. 231*, 459−474.
Yang, A.-S., Gunner, M. R., Sampogna, R., Sharp, K., & Honig, B. (1993) *Proteins: Struct., Funct., Genet. 15*, 252−265.
You, T. J., & Bashford, D. (1995) *Biophys. J. 69*, 1721−1733.

BI9601565