PROTEINS: Structure, Function, and Genetics 15:252–265 (1993)

On the Calculation of pK_as in Proteins

An-Suei Yang, M. R. Gunner, Rosemary Sampogna, Kim Sharp, and Barry Honig
Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York 10032

ABSTRACT **This paper describes a general method to calculate the pK_as of ionizable groups in proteins. Electrostatic calculations are carried out using the finite difference Poisson–Boltzmann (FDPB) method. A formal treatment of the calculation of pK_as within the framework of the FDPB method is presented. The major change with respect to previous work is the specific incorporation of the complete charge distribution of both the neutral and charged forms of each ionizable group into the formalism. This is extremely important for the treatment of salt bridges. A hybrid statistical mechanical/Tanford–Roxby method, which is found to be significantly faster than previous treatments, is also introduced. This simplifies the problem of summing over the large number of possible ionization states for a complex polyion. Applications to BPTI and serine proteases suggest that the calculations can be quite reliable. However, the necessity of including bound waters in the treatment of the Asp-70 . . . His-31 salt bridge in T4 lysozyme and experience with other proteins suggest that additional factors ultimately need to be considered in a comprehensive treatment of pK_as in proteins.**

Key words: titration curves, electrostatics, ionization, ion pairs, salt bridges

INTRODUCTION

Solution pH is a major determinant of protein stability as well as of protein function. For this reason, understanding the factors that govern the pK_as of amino acids in proteins has been a classical problem in biophysical chemistry.[1] In recent years, renewed interest in this problem has been due to both theoretical advances in the treatment of the electrostatic properties of proteins[2,3] as well as to improved experimental probes of electrostatic effects on protein stability (ref. 4 and references therein). The basic theoretical problem in calculating pK_a values is to predict how the protein environment shifts the pK_a of an amino acid in the protein from that of the isolated amino acid in solution. The underlying factors affecting these shifts are well understood, at least in a qualitative sense. For instance, the equilibrium between the ionized and neutral states in aqueous solution is determined in part by the different solvation free energies of the two species. In a protein, although interactions with water are restricted, they are replaced by new interactions with the electrons, permanent dipoles of the protein, and other titratable groups. These factors will be discussed in greater detail below.

Early treatments of pK_a shifts[5,6] assumed that the intrinsic pK_a of a group in the protein (defined here as the pK_a of an amino acid in a state where all other amino acids are neutral) is identical to that of the isolated amino acid in solution. All shifts were then attributed to electrostatic interactions among the different ionized groups which were calculated with Tanford–Kirkwood theory.[6] This essentially is a solution to the Poisson–Boltzmann equation for charges placed in a sphere of lower dielectric constant than the surrounding solvent. In order to determine the titration curve of a protein, it is necessary to calculate simultaneously the ionization state of each ionizable group at every pH. This complicated problem was solved initially by Tanford and Roxby using an iteration scheme discussed further below. However, the Tanford–Roxby method, as originally introduced, encountered difficulties in reproducing experimentally determined titration curves. In an attempt to improve the applicability of the method, Gurd and co-workers introduced a correction term based on the accessibility of the amino acid to water.[7,8] They retained the assumption that intrinsic pK_as are essentially unmodified from those in solution, but allowed an empirical adjustment of $\pm$ 0.5 pK_a units.[7]

A fundamental problem with any version of the Tanford–Kirkwood method, as demonstrated most convincingly in the studies of Warshel and co-workers,[9,10] is the assumption that intrinsic pK_as are unmodified in proteins. In many cases, the free energy changes resulting from the loss of aqueous solvation and from the effects of permanent and induced dipoles in the protein outweigh the electrostatic interactions between different ionizable groups. Warshel and co-workers carried out the first

Received May 26, 1992; revision accepted July 27, 1992.
Address reprint requests to Dr. Barry Honig, Department of Biochemistry and Molecular Biophysics, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, New York, NY 10032.
Current address of M. R. Gunner: Department of Physics, City College of New York, New York, NY 10031.
Current address of Kim Sharp: Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA 19104.

detailed characterization of all of the factors that determine intrinsic pK_as. A simplified treatment of interactions between ionizable groups also was included, but ionic strength effects were not considered.

Advances in numerical methods now make it possible to solve the Poisson–Boltzmann (PB) equation for molecules of arbitrary shape and charge distribution.[2] We use the term FDPB method to describe finite difference solutions to the PB equation. This approach provides a detailed description of the full range of electrostatic interactions including those due to permanent and induced dipoles in the protein, the dielectric effects of the solvent and the effects of ionic strength. An early application to the pK_a shifts resulting from interactions between ionizable amino acids yielded good agreement with the effects detected with site-directed mutagenesis.[11,12] The first full pK_a calculation using the FDPB method was recently reported by Bashford and Karplus[13] in their study of lysozyme. The various ionization states that arise (see below) were treated using statistical mechanics. This was augmented by a reduced site approximation which limited the number of ionization states that need be considered by assuming that amino acids whose pK_as are far from a particular pH are exclusively in the predominant ionization state. More recently, Beroza et al.[14] have used a Monte Carlo approach to deal with the problem arising from multiple titrating sites.

The present work is similar to that of Bashford and Karplus[13] and Beroza et al.,[14] except that we have dropped the approximation that the ionized state be represented by a charge located on a single atom and that the unionized state is completely nonpolar. These approximations can lead to errors when ion pairs are considered since the energy of the state that corresponds to one group charged and the other neutral clearly will be sensitive to the detailed charge distribution assumed for each group. In the calculations described below, interactions between two members of a salt bridge in both charged and neutral forms are included explicitly in the pK_a calculations, requiring a change in the expressions that were introduced in previous work. Several strategies are considered for treating the multiple site combinatorial problem and a new method that combines the Tanford–Roxby iteration scheme for weakly interacting groups and treats the remaining interactions with the standard statistical mechanical expression is introduced. Comparisons with experimental results are made for BPTI, several serine proteases and T4 lysozyme.

METHODS

General Formalism

Single titratable sites

We first consider a protein with a single titratable site and compare its pK_a to that of an isolated amino acid. The fractional degree of protonation, f_i, of any ionizable group, i, is given by

$$\ln[f_i/(1 - f_i)] = 2.3(\mathrm{p}K_{ai} - \mathrm{pH}). \quad (1)$$

If the group corresponds to an isolated amino acid in solution, its pK_a is written as pK_{ai}^0. The pK_{ai}^0 values used were Asp 3.9, Glu 4.3, Lys 10.5, Arg 12.5, His 6.5, C-terminus 3.6, and N-terminus 8.0. If the group is in a protein with a single site, its pK_a is referred to as pK_{ai}^{int}, the "intrinsic" pK_a, which is given by

$$\mathrm{p}K_{ai}^{int} = \mathrm{p}K_{ai}^0 - \gamma(i)\Delta\Delta G_i^{env}/2.3kT. \quad (2)$$

$\Delta\Delta G_i^{env}$ is the change in electrostatic energy of ionizing the group in the protein environment relative to solution. $\gamma(i) = -1$ or 1 for an acidic or basic group, respectively. Referring to the thermodynamic cycle in Figure 1 (see also refs. 2 and 13), $\Delta\Delta G_i^{env}$ is given by

$$\Delta\Delta G_i^{env} = -\gamma(i)\{\Delta G_i^{env}(A) - \Delta G_i^{env}(AH)\} \quad (3)$$

where $\Delta G_i^{env}(A)$ is the difference in the free energy of interaction of A (in the unprotonated state) with its protein environment relative to the reference state in solution. $\Delta G_i^{env}(AH)$ is the corresponding term for the protonated state. Each of the terms in Eq. (3) can be further broken up into two contributions

$$\Delta G_i^{env} = \Delta G_i^{rf} + \Delta G_i^{perm} \quad (4)$$

where ΔG_i^{rf} is the difference in reaction field energy in the isolated state and in the protein, and ΔG_i^{perm} is the contribution of the permanent dipoles and any other pH independent electric field in the protein. The latter would include, for example, the effects of bound ions and water molecules. The reaction field term is just the contribution of those parts of the system treated as a dielectric continuum. In FDPB calculations, the reaction field in the aqueous phase yields the solvation energy. In the protein it arises from (the now reduced) interactions with the solvent as well, but also includes the effects of the polarizable electrons and minor rearrangements of polar groups which are incorporated through the use of a dielectric constant of 4.[15] It is evident that the ionization state of a group in a protein will always be

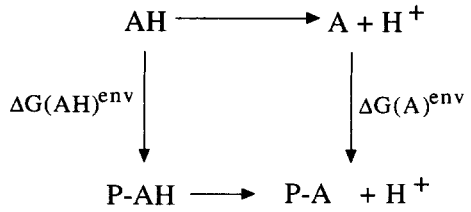

Fig. 1. Deprotonation of an acidic group in solution and in a protein. $\Delta G(AH)$ and $\Delta G(A)$ are the differences in the electrostatic free energy of AH and A in the protein and solution, respectively.

shifted by the reaction field term to favor the neutral species. This effect increases as the accessibility to water decreases, but favorable interactions of the ionized state with dipolar groups can compensate for this.

Multiple sites

When the protein has more than one titratable group, the mutual charge–charge interactions at the titrating site become pH dependent. pK_{ai} is now obtained by determining the pH at which the group i is 50% protonated. This is done by calculating a titration curve for the entire system. In most cases, the pK_a of group i can now be written in the form

$$pK_{ai} = pK_{ai}^{int} + \Delta pK_{ai}^{ttr}. \quad (5)$$

The last term in Eq. (5) is simply due to charge–charge interactions. Under certain circumstances, the titration curve of site i fails to follow Eq. (1), and Eq. (5) does not hold. In these cases, there may not be a unique pH at which a group is 50% ionized (see, e.g., Fig. 4). This leads to a breakdown of the Tanford–Roxby[5] iteration scheme as will be discussed further below.

The titration curve of the system can be obtained from a statistical mechanical average over the 2^N states that arise from the N ionizable residues. A given ionization state of the protein, n, where $n = 1$ to 2^N, can be defined in terms of the vector $\delta_n(i)$, $i = 1$ to N, where $\delta_n(i)$ is 0 when the group i is neutral and 1 when it is charged. (Note that even when a group is neutral, its atoms bear partial charge.) The net charge, Q_n, on the protein in state n is given by

$$Q_n = \sum_{i=1}^{N} q_n(i) = \sum_{i=1}^{N} \delta_n(i)\,\gamma(i) \quad (6)$$

where $q_n(i)$ is the charge on group i in state n and $\gamma(i) = -1$ or 1 for an acidic or basic group i, respectively. A reference state of zero free energy is defined as corresponding to all ionizable groups in their neutral form [i.e., $q_n(i) = 0$ for all i]. ΔG^n is the free energy of the nth state and is given by

$$\Delta G^n = \sum_{i=1}^{N} \{q_n(i)[2.3kT(\mathrm{pH} - pK_{a_i}^{int})] + \delta_n(i) \sum_{1 \le j < i} \delta_n(j)\Delta G^{ij}\}. \quad (7)$$

By writing out expressions for the free energy at each step in the process of charging both sites i and j from the neutral form, it can be verified that

$$\Delta G^{ij} = \Delta G^{ij}(1,1) - \Delta G^{ij}(1,0) - \Delta G^{ij}(0,1) + \Delta G^{ij}(0,0). \quad (8)$$

$\Delta G^{ij}[\delta(i), \delta(j)]$ is the interaction free energy in kcal/mol between amino acids i and j in their charge states defined by $\delta(i)$ and $\delta(j)$. Thus, for example, $\Delta G^{ij}(1,1)$ is the interaction energy between amino acids i and j when both are charged. Four terms arise in the last sum in Eq. (7) because the energies of the states corresponding to one amino acid charged and one neutral appear in the pK_a^{int} term. In previous work where charged groups were treated as point charges and neutral groups as having zero charge, ΔG^{ij} was simply the electrostatic interaction energy between point charges.[13,14] A recent study reports the use of distributed charges although an expression equivalent to Eq. (8) was not provided.[16]

The average charge, $<\rho_i>$, of each group is determined from the statistical mechanical average

$$<\rho_i> = \sum_{n=1}^{2N} \delta_n(i)\gamma(i)\exp(-\Delta G^n/kT)/Z \quad (9a)$$

where Z is the partition function of the system

$$Z = \sum_{n=1}^{2N} \exp(-\Delta G^n/kT). \quad (9b)$$

The net average charge, Q, on the protein is just the sum of the individual $<\rho_i>$s.

Calculation of Titration Curves

Equations (7)–(9) provide a full statistical mechanical prescription for calculating the titration curve of N ionizable residues. However, the large number of possible states makes this treatment impractical as N reaches several tens of titratable groups. The first attempt to circumvent the problem was introduced by Tanford and Roxby[5] who assumed that each residue could be defined in terms of an average protonation state based on interresidue interactions (see below). In this approximation, the average charge of the groups is assumed to be proportional to the mean potential between each pair of ionizable groups. This approximation is generally valid but breaks down when two groups have similar pK_as (less than a 1 pK_a unit difference) and the interaction between them is on the order of 1 pK_a unit or more[17] (see also below). Bashford and Karplus[17] used the full statistical mechanical expression, but limited the number of states considered by introducing a reduced site approximation in which residues whose intrinsic pK_as are far from the pH of interest were assumed to be in their appropriate protonation state. More recently, Beroza et al.[14] used a Monte Carlo approach to deal with the problem of sampling the large number of possible states. Here we introduce a modified Monte Carlo method as well as a new approach which exploits the Tanford–Roxby approximation for weakly interacting groups and uses the full statistical mechanical summation for strong interactions.

Tanford–Roxby approximation

The Tanford–Roxby approximation[5] assumes the average charge of each residue depends upon the average charges of all other residues. That is, it

avoids the time-consuming summation over 2^N protonation states in Eq. (9) by using fractional charges which can be interpreted as averages over protonation states. (The relationship between the Tanford–Roxby approximation and the full statistical mechanical treatment is clearly developed in a recent paper by Bashford and Karplus.[17]) In this method, each site is assigned a specific charge. These charges are included in an iteration scheme where they are updated at each step until a self consistent solution is obtained. The method can lead to errors,[17] but is quite fast and, in many cases, as discussed further below, quite accurate.

To start the Tanford–Roxby iteration, an initial guess for each fractional charge is made. Typically, this is set according to the intrinsic pK_a of each group and the bulk pH using Eq. (1). (Note that the average charge of a basic group is just f_i while it is (f_i − 1) for an acidic group.) At each step, t, in the iteration the pK_a of the group is updated with the equation

$$pK_{a_i}(t) = pK_{a_i}^{\text{int}} - \sum_{j=1, j\neq i}^{N} \gamma(i) \mid <\rho_j> \mid \Delta G^{ij}/(2.3kT) \tag{10}$$

where the pairwise charge–charge interaction term, ΔG^{ij}, was introduced in Eq. (7). The fractional charge, $<\rho_j>$, on each group is recalculated with Eq. (1) at each step of the iteration using the new pK_{aj} value until they converge to a set of self-consistent solutions. Generally fewer than 5 cycles are required. This is usually very fast and takes only a few seconds of CONVEX C2 CPU time. Since the Tanford–Roxby algorithm uses Eq. (1) to convert pK_{ai} to $<\rho_i>$, the method breaks down when titration curves have significantly different shapes than the form shown in Eq. (1). This occurs when the ionizable groups are strongly coupled and results in the Tanford–Roxby iteration not converging or converging to incorrect values.

Hybrid statistical mechanical/Tanford–Roxby

Our hybrid statistical mechanical/Tanford–Roxby procedure relies on the fact that charge–charge interactions in a protein fall off rapidly with distance[18] so that interresidue interactions can be expected to be quite small beyond some cutoff distance. Residues separated by greater than this cutoff (here assumed to be 7 Å) are treated with the Tanford–Roxby approximation while residues falling within the cutoff distance are treated with the statistical mechanical summation. For the small globular proteins we have studied, the number of ionizable residues within 7 Å of a particular residue does not exceed 5. (An equivalent approach would be to use an energetic rather than a distance cutoff as a basis for dividing residues into different groups.)

Consider a protein with N ionizable groups. Assume that all groups designated by 1 to j − 1 are within 7 Å of a particular group, s, under consideration, while those designated by j to N are outside the cutoff radius. In the first step of the procedure, the average charge of each group in the protein is calculated with the Tanford–Roxby procedure. The average charge of groups more than 7 Å away from group s is then kept fixed for the remainder of the calculation on s. Only groups within 7 Å are treated with the statistical mechanical expressions [Eqs. (7) and (9)].

Equation (7) now takes the form

$$\Delta G_n = \sum_{i=1}^{j-1} \{q_n(i)[2.3kT(\text{pH} - pK_{a_i}^{\text{int}})] + \delta_n(i) \sum_{k=j}^{N} |<\rho_k>|\Delta G^{ik} + \delta_n(i) \sum_{1\leq m<i} \delta_n(m)\Delta G^{im}\}. \tag{11}$$

The first term of Eq. (11) shows that each residue s will have 2^{j-1} states associated with it, where j − 1 denotes the number of residues within 7 Å of s. The second term corresponds to the interaction of the charge on group i with the fixed charges on groups located more than 7 Å away. The third term describes the interaction of group i with groups located closer than 7 Å. Note that the number of groups in the sets k and m depends on the identity of group s and that ΔG^n is pH dependent. The average charge, $<\rho_s>$, on group s can now be calculated from Eq. (9). The sum over 2^N states in Eq. (9) is now replaced by a sum over 2^{j-1} states, where j − 1 is usually less than 5.

The new average charge, calculated as described above, is assigned to group s. The same procedure is now applied to a second group and continues for residues 1 to N, repeating until a self-consistent set of average charges is obtained at each pH. It is worth noting that a cutoff radius of 0 Å results in the Tanford–Roxby procedure while a cutoff radius of infinity yields the full statistical mechanical expression. The cutoff of 7 Å used in this work was found to provide a satisfactory balance between speed and reliability. A somewhat larger cutoff might be required for membrane proteins where groups are removed from water, hence, values of ΔG^{ij} are larger.

Monte Carlo method

The occupancies of the charged and neutral acids and bases also can be obtained by a Monte Carlo sampling technique. Each step starts by switching the ionization state of a randomly chosen titratable residue. As described by Beroza et al.,[14] if sites are closely coupled when only single flips are allowed, the system may be trapped in a local minimum. This can be avoided if both sites are permitted to change ionization state in a single Monte Carlo step, a procedure that essentially provides for intrasystem proton transfer. Single changes are still required to provide for net proton uptake or release. The imple-

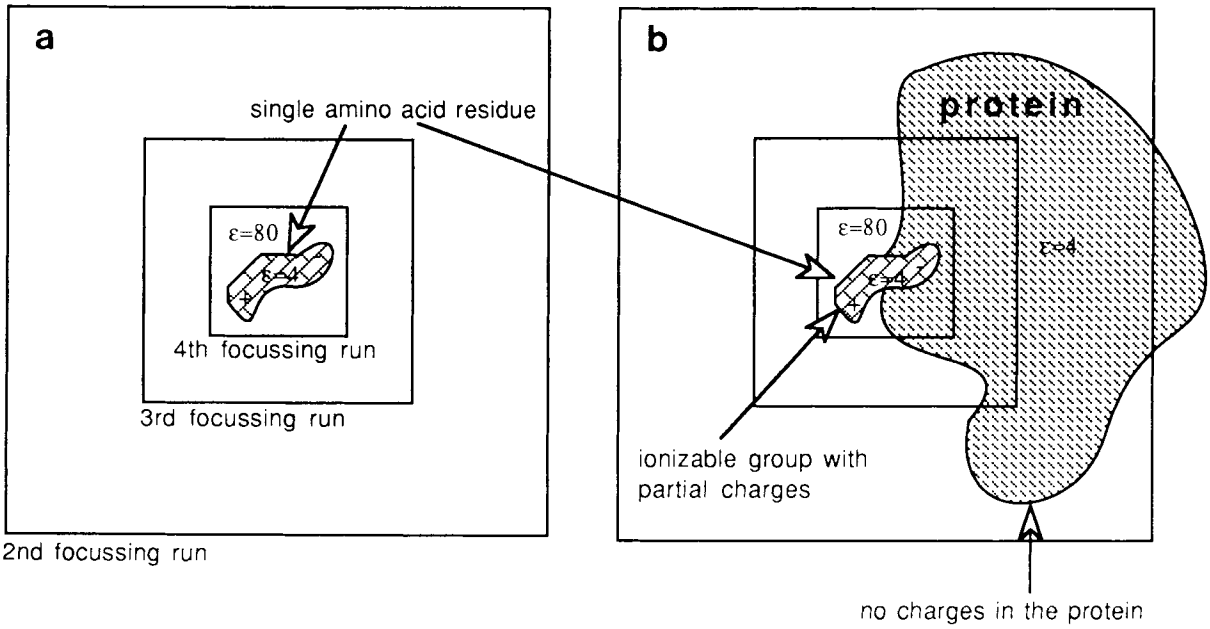

Fig. 2. Calculation of the reaction field term, $\Delta\Phi^{rf}$, (**a**) in water and (**b**) in a protein. $\Delta\phi^{rf} = \phi^{rf}$(protein) − ϕ^{rf}(water). (Note that the first focussing run is not shown.)

mented procedure makes two flips 50% of the time. The second site is chosen from a list of groups interacting with the first site by greater than 1.4 kcal/mol. The free energy of the resultant state, obtained by changing the ionization state of either one or two sites, is calculated and judged by Metropolis sampling. If the change is accepted then this Monte Carlo step is concluded and the new state is now used to initiate the next step.

Standard runs begin with the output of a Tanford–Roxby calculation with site occupancies rounded off to zero or one. The system is equilibrated by an initial phase of Monte Carlo sampling carried out without saving the results for approximately 10% of the total steps in the calculations. Next, a running average of the occupancy of accepted states provides the Boltzmann distribution of ionization states. For each protein, convergence is checked by comparing the results of runs with randomly generated starting states and by monitoring the average occupancy of sites after different numbers of steps. The rate of change of occupancy decreases as the number of Monte Carlo steps increases. At convergence the change in occupancy fluctuates around zero.

Electrostatic Calculations

The finite difference Poisson–Boltzmann (FDPB) method

Details of the FDPB method have been reviewed in a recent publication.[2] Briefly, the protein is treated as a low dielectric cavity (dielectric constant = 4) embedded in aqueous medium of dielectric constant 80 with an electrolyte behaving according to the Poisson–Boltzmann equation. Hydrogen atoms are built onto the protein prior to electrostatic calculations using the DISCOVER program (Biosym Technologies Inc.). Separate structures are constructed for the charged and neutral configurations. Charges are placed at the centers of the atoms (the DISCOVER charge set was used in this work).[19] The interior of the protein is defined by the solvent accessible surface using a probe of radius 1.4 Å. Salt is excluded at distances less than 2 Å from the surface. The radii assigned to different atoms were carbon 1.9 Å, nitrogen 1.65 Å, oxygen 1.6 Å, hydrogen 1.0 Å, and sulfur 1.9 Å. The entire system is mapped onto a lattice of dimension 65 × 65 × 65. In this work, a salt concentration of 0.14 M (Debye length = 8.0 Å) was used throughout. The Poisson–Boltzmann equation was solved with the enhanced DelPhi program.[20]

Reaction field (desolvation) energies

ΔG^{rf} [Eq. (4)] is determined from the difference in the reaction field potential, $\Delta\Phi^{rf}$, at a charged atom i in an isolated amino acid side chain, which is assumed to have the same structure as the corresponding side chain in the protein (see Fig. 2).

$$\Delta G^{rf} = \Sigma q_i \Delta\Phi^{rf}/2. \quad (12)$$

The greatest precision problems involved in using the FDPB method are encountered in the calculation of these terms. In general, it is necessary to test whether results are dependent on grid size. The following procedure, as shown in Figure 2, is quite reliable.

The protein is mapped onto a lattice with a reso-

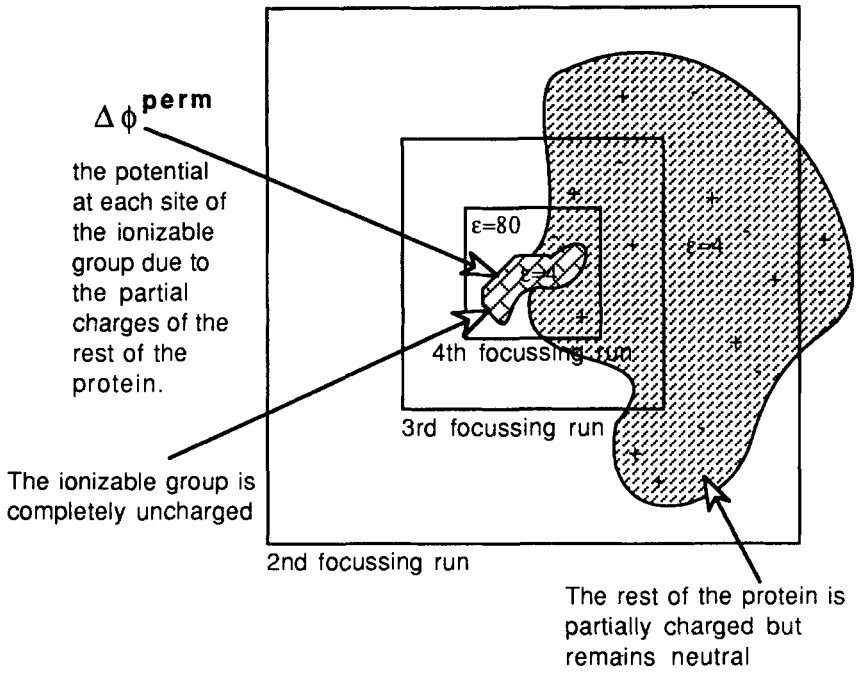

Fig. 3. Calculation of the permanent dipole term, $\Delta\Phi^{perm}$.

lution of about 0.7 grid per Å. The center of the residue is placed at the center of the lattice. The reaction field at the charge site i is obtained following four successive focussing runs,[21] at which point the resolution has increased to about 3 grids per Å.

Permanent dipoles

ΔG^{perm} is the free energy term arising from interactions between a particular ionizable amino acid and the potential, $\Delta\Phi^{perm}$, due to partially charged atoms in the rest of the protein. This includes partial charges on backbone atoms, water molecules, and sidechains in their neutral forms (see Fig. 3).

$$\Delta G^{perm} = \Sigma q_i \Delta\Phi^{perm} \quad (13)$$

where the sum is over all charged sites in the side chain of the ionizable group. C- and N-terminal groups are treated as individually titrating sites and separate from their side chain atoms. The center of the ionizable group in question is then placed at the center the lattice and four focussing runs, to about 3 grid per Å, are carried out.

Interactions between charged groups

The calculation of ΔG^{ij} is equivalent to that of ΔG^{perm}, except that only two groups, i and j, are involved. The rest of the protein, including the backbone atoms of the groups in question, is assumed to be uncharged. We have found that the last three terms in Eq. (8), which involve neutral forms of i or j, are significant only when groups i and j are in close proximity (less than 3.5 Å) to each other. In all other cases they have been assumed to be zero for computational efficiency.

Coordinates

Coordinates for BPTI,[22,23] bovine trypsin,[22,24] chymotrypsin,[25,26] elastase,[27,28] and T4 lysozyme[29,30] were obtained from the Protein Data Bank[31] at Brookhaven National Laboratory.

RESULTS

Comparison of Methods

In general, the Tanford–Roxby approximation works quite well for weakly interacting groups but breaks down when groups with similar pK_as interact strongly (see discussion in Bashford and Karplus[17]). Figure 4 shows a group of titrations where two interacting acids have identical pK_a^{int}. The initial portion of the titration curve, where only a small fraction of the sites is ionized, is relatively independent of the interaction energy between the sites (ΔG^{ij}). However, as ΔG^{ij} increases, the fraction ionized increases more slowly, effectively buffering the system from changes in the environment. The titration curves show that this coupled system can be characterized as having two effective pK_a values with half of each site titrating at $pK_a^{int} - \log(0.5)$ and the other half at $pK_a^{int} + \Delta G^{ij}/1.36 + \log(0.5)$.[1] When two sites have different intrinsic pK_as, extended regions with little change in ionization state are also produced, but with different occupancies. For example, two groups differing in intrinsic pK_a by one unit will exhibit behavior where one is 10%

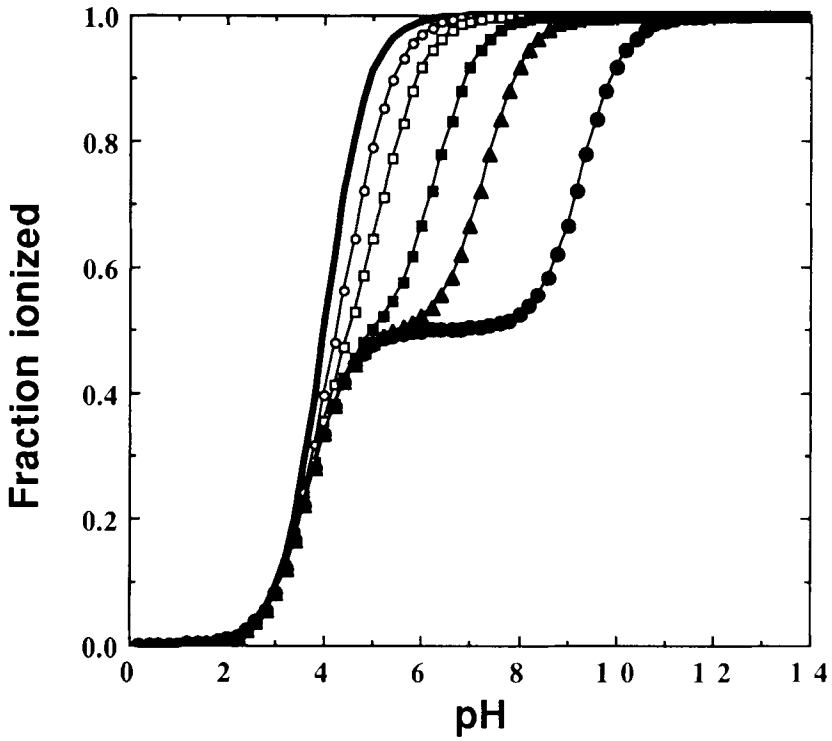

Fig. 4. Titration curves calculated by the statistical method for two acids each with $pK_a^{int} = 4$. The bold line shows the results for noninteracting sites; the other lines show the results with interaction energies of (○) 0.69, (□) 1.38, (■) 2.76, (▲) 4.14, (●) 6.9 kcal/mol. Since pK_a^{int} is the same for both sites the titration curve for the two sites are superimposed.

ionized and one 90% ionized over a pH region of approximately $\Delta G_{ij}/1.36$ pH units (results not shown).

Figure 5 compares the results obtained from the Tanford–Roxby approximation, and for the two flip Monte Carlo scheme, for one of the cases shown in Figure 4 (ΔG^{ij} = 2 pH units = 2.76 kcal/mol at 25°C). Under these conditions, the Tanford–Roxby calculation becomes unstable. As seen in the figure, regions where the two sites behave identically are interrupted by a region in the center of the titration where their behavior is predicted incorrectly. Note however, that the method overestimates the charge of one site and underestimates the charge of the other. Thus, the errors for the total charge on the system are much smaller than those found when the fractional ionization of individual sites are determined[17] (see Fig. 5b). The two flip Monte Carlo scheme produces results in excellent agreement with the exact statistical mechanical expression, while a Monte Carlo procedure allowing for only single steps is seen to converge far more slowly.

The five available methods [Tanford–Roxby, hybrid statistical mechanical/Tanford–Roxby, Monte Carlo, reduced (Bashford and Karplus),[17] and full statistical mechanical] were tested on a number of proteins (BPTI, apo-myoglobin, and bacteriorhodopsin). All five methods produced essentially identical results for BPTI with differences of less than 0.1%. BPTI has 16 ionizable sites besides tyrosine (see Table I). The Tanford–Roxby iteration took less than 1 second CPU time on a Convex C2, the hybrid method took 2 sec, the Monte Carlo method took 2 min, the reduced site statistical mechanical method[17] took 3 sec, while the full statistical mechanical treatment required 3 hr. All timings for different methods in this paper are for a complete protein titration curve. This involved a total of 70 separate sets of calculations, one for each pH between 0.2 and 14, at intervals of 0.2 pH units. The Tanford–Roxby, hybrid, and Monte Carlo methods also produced identical results for apo-myoglobin (58 ionizable groups excluding tyrosine) and took 5 sec, 20 sec, and 7 min, respectively. The reduced and full statistical mechanical procedures were not run due to their large computational demands.

It should be pointed out that the computational time required by the reduced statistical and full statistical mechanical methods increases exponentially with increasing number of ionizable groups, whereas the hybrid and Monte Carlo methods increase linearly with N. For bacteriorhodopsin (considering 28 ionizable sites), the hybrid and Monte Carlo methods yielded the same results, which were significantly different from those produced by the Tanford–Roxby iteration. The latter method breaks down

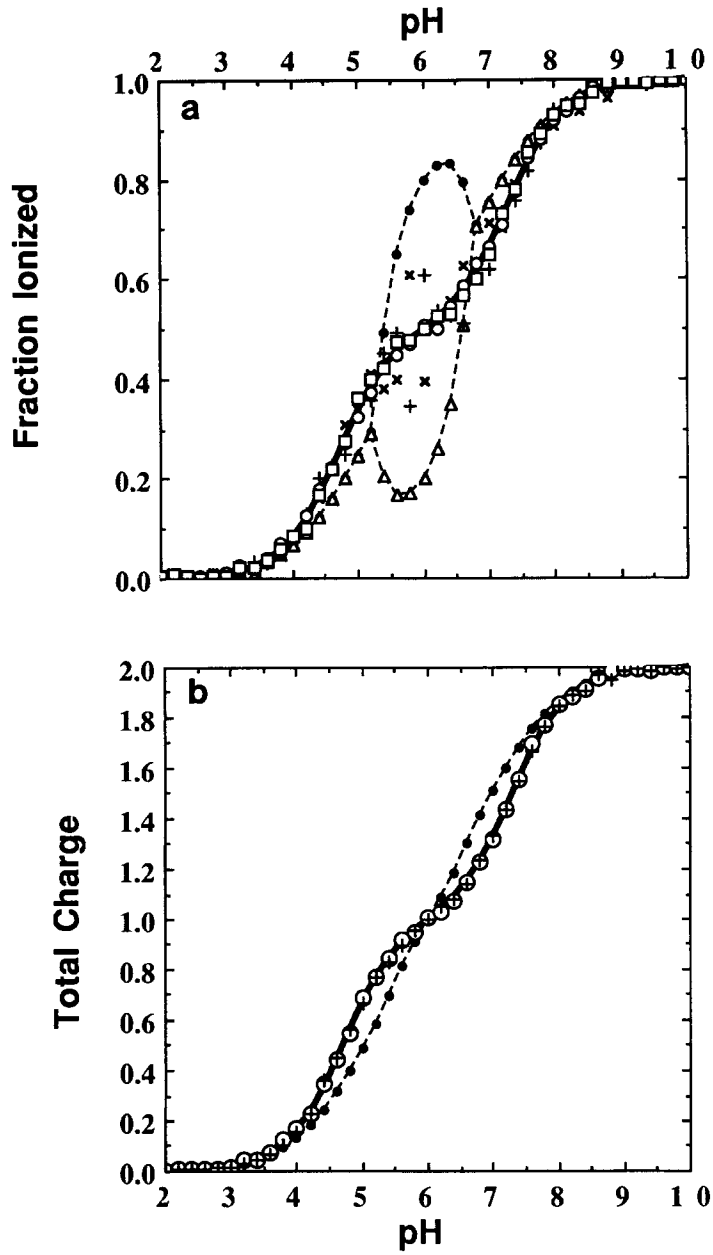

Fig. 5. Comparison of four methods for calculating the titration of two acids with $pK_a^{int} = 4$. The interaction between the sites is 2.76 kcal/mol. (a) The fraction of each site that is ionized. The dark solid line is obtained with complete statistical method. The ionization state of the two sites are the same at all pHs; Monte Carlo search where a single flip is allowed on each step (+, ×) or with 2 flips are allowed 33% of the time for strongly interacting sites (○, □). One thousand Monte Carlo steps were taken. The dotted lines represent the Tanford–Roxby calculation (●, △). (b) The total charge on the system calculated with the statistical method (solid line), single flip Monte Carlo method (+), 2 flip Monte Carlo method (○), and Tanford–Roxby (dotted line).

when strong interactions between ionizable groups occur, as expected in the interior of a membrane protein. For the proteins we studied, the hybrid method was found to be at least an order of magnitude faster than the Monte Carlo method. It should be emphasized that the CPU times reported here are only for determination of the Boltzman distribution of ionized states. Extensive FDPB calculations must be first carried out to provide the input for the free energy determinations.

BPTI

BPTI has a total of 16 ionizable residues excluding tyrosines. Of the 120 possible pairs of interacting residues, about 80% had interaction energies between ionized forms (ΔG^{ij}) less than 0.1 kcal/mol.

TABLE I. pK_as in BPTI

Residue	pK_a calculated	pK_a experimental[32]
NTE-1	7.0	8.1[36]
ARG-1	13.8	
ASP-3	3.6	3.0[42]
GLU-7	3.4	3.7[42]
LYS-15	10.7	10.6[43]
ARG-17	12.5	
ARG-20	13.2	
LYS-26	10.8	10.6[43]
ARG-39	12.4	
LYS-41	10.3	10.8[43]
ARG-42	12.9	
LYS-46	10.3	10.6[43]
GLU-49	4.5	3.8[42]
ASP-50	1.7	3.4[42]
ARG-53	13.6	
CTE-58	3.6	2.9[44]

Only one interaction, the salt bridge between Asp-50 and Arg-53, was greater than 1 kcal/mol.

The pK_as of all titratable groups in BPTI except arginines have been determined by NMR.[32] These are shown in Table I together with our calculated values. The agreement between theory and experiment is quite good (within ~1 pK_a unit) except for Asp-50 and the N-terminus. The calculated pK_a of Asp-50 is lowered due to its strong interaction with Arg-53. However, it should be pointed out that recent NMR and crystallographic evidence suggests that the salt bridge Asp-50 . . . Arg-53 found in the crystal may not be present in solution.[33–35] Thus, it is possible that the discrepancy is due to our use of a particular crystal structure rather than to errors in the parameters or theory.

Evidence from NMR experiments also reveal a salt bridge interaction between the amino and carboxyl termini in BPTI.[36] This salt bridge conformation was not seen in the crystal structure used here where the N- and C-termini are 7 Å apart. This specific salt bridge interaction is estimated to be 1 kcal/mol[32] (which is higher than the interaction energy we calculated for this pair). Using this value in our calculations further shifts the pK_as of the N- and C-termini from 7.0 and 3.6 (see Table 1) to 7.8 and 2.8, respectively (at 25°C). These pK_a values are quite close to the NMR determined ones, indicating that differences in protein conformation in the crystal and in solution may be responsible for the discrepancy between experiment and theory in Table I. Thus, the calculations can highlight differences between the structures of the protein in solution, where the pK_a determinations are made, and in the crystal, which provides the X-ray coordinates.

Figure 6 plots values of $\Delta\Delta G^{rf}$ and $\Delta\Delta G^{perm}$ for each of the ionizable groups in BPTI. $\Delta\Delta G^{rf}$ is positive for each residue and its magnitude indicates the desolvation penalty associated with its presence in the protein. $\Delta\Delta G^{perm}$ is negative in almost every case. The obvious pattern, which has been pointed out previously by Russell and Warshel in their study of this system,[9] is that the permanent dipoles on the protein tend to compensate for the loss of aqueous solvation. This is the reason intrinsic pK_as are often close to those of the isolated amino acids. When this is the case, the actual pK_as will also be relatively unshifted, unless, of course, there are strong interactions between different ionizable groups.

Serine Proteases

We also studied pK_as in four serine proteases. Of special interest were the pK_as of His-57 and Asp-102 which are part of the active site catalytic triad. These groups form an ion pair in the active state of serine proteases and their mutual interaction is an important component of the catalytic mechanism. In particular, the negative charge on Asp-102 stabilizes the positive charge on His-57 in the transition state.[37,38] For all the proteases, Asp-102 is completely buried and has a pK_a below 2 while His-57 is partially buried and has a pK_a of 6.8.[39]

A complication arises in the treatment of ion pairs since coordinates are available for only one protonation state of the pair. In this case, it is the state where Asp-102 is charged and His-57 is neutral. Thus, it is necessary to decide to which oxygen the hydrogen in the neutral form of Asp-102 is attached. We considered both possibilities and the results obtained in either case were found to be nearly identical. We chose the more likely neutral Asp-102 configuration in which the hydrogen is attached to the oxygen which does not form a hydrogen bond with His-57.

Figures 7a and b plot $\Delta\Delta G^{rf}$ and $\Delta\Delta G^{perm}$ of His-57 (Fig. 7a) and Asp-102 (Fig. 7b) for the four different serine proteases. The figures also plot ΔG^{W}, which is the free energy contribution resulting from the sum of the charge–charge interactions of a titratable residue with all other ionizable groups in the protein at a particular pH. ΔG^{W} is given by

$$\Delta G^{W} = \Sigma|<\rho_k>|\Delta G^{ik} \qquad (14)$$

where the index i represents the group being considered (in this case either His-57 or Asp-102) and k is the index for all ionizable residues other than i. $<\rho_k>$ is the average charge of group k at a given pH. Values in these figures represent results obtained from calculations done at pH 6. The term ΔG^{W} is most heavily biased by the favorable electrostatic interaction between His-57 and Asp-102, more distal residues making only a minor contribution. These charge-charge interactions largely compensate for the partial burial of His-57 ($\Delta\Delta G^{rf}$ = ~4 kcal/mol), leading to a pK_a that is relatively unshifted from an isolated histidine. The calculated

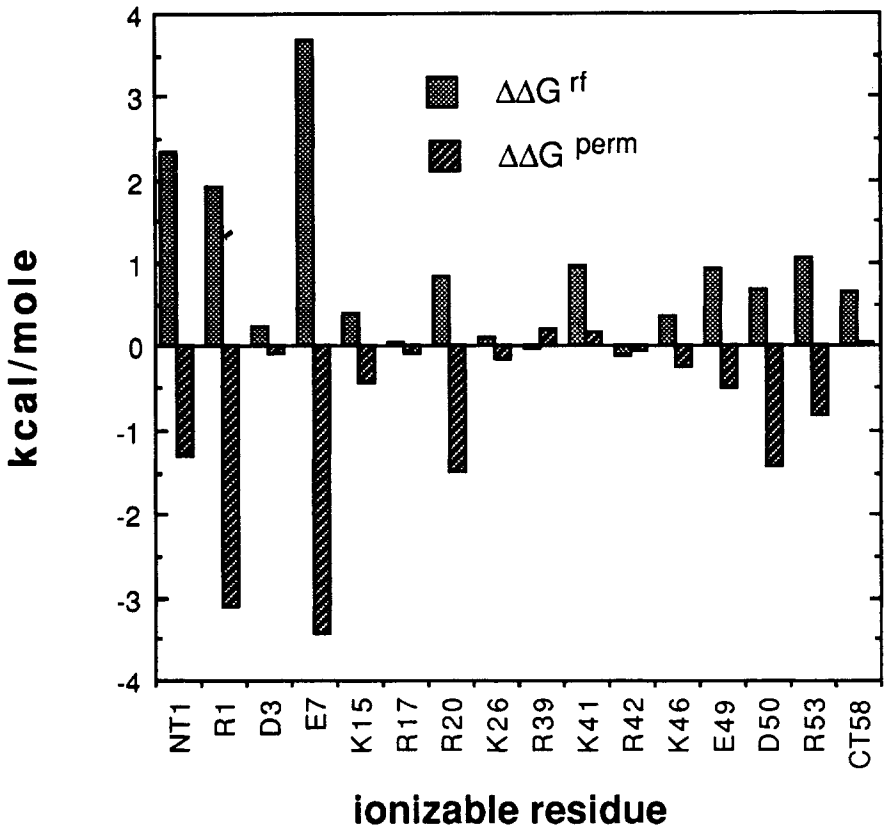

Fig. 6. Calculated free energy contributions, $\Delta\Delta G^{rf}$, $\Delta\Delta G^{perm}$, to pK_as in BPTI. Salt concentration 0.14 M.

pK_as for His-57 in the catalytic triad of rat trypsin, bovine trypsin, chymotrypsin, and elastase are 6.4, 7.1, 6.8, and 7.0, respectively.

The low pK_a of Asp-102 is somewhat surprising since it is totally buried ($\Delta\Delta G^{rf}$ = ~9 kcal/mole). However, its charged carboxyl group forms four hydrogen bonds with neighboring residues leading to a large negative value for $\Delta\Delta G^{perm}$. This, in addition to electrostatic interactions with other ionizable residues (mostly His-57), actually produce a pK_a a little less than zero (−0.6, −1.2, −1.7, −1.4 for rat trypsin, bovine trypsin, chymotrypsin, and elastase, respectively), in agreement with the experimental observation of a pK_a less than 2. When the same calculations were run using only a single charged atom for the ionized state,[13] pK_a values for His-57 and Asp-102 of bovine trypsin were 8.4 and −4.3, respectively. This illustrates the problem of using a simplified charge distribution when ion pairs are included.

T4 Lysozyme

The pK_as of Asp-70 and His-31 of T4 lysozyme have been studied using site directed mutagenesis and NMR techniques.[40] In the native protein, the apparent pK_as of His-31 and Asp-70 are 9.1 and 1.4, respectively. The large shifts relative to the pK_as of the isolated amino acids appear due to the salt bridge formed between these two residues. This contrasts with the similar salt bridge in serine proteases (see above) which has only a marginal effect on the pK_a of His-57. The important differences between the two cases are the polarity of the immediate neighboring residues and the solvent accessibility of the salt bridges. Although Asp-102 in the serine proteases is buried under the solvent accessible surface of the protein, its charged carboxyl group is stabilized by four hydrogen bonds from neighboring residues in addition to its interaction with His-57. In contrast, the His-31 . . . Asp-70 salt bridge in T4 lysozyme is on the surface of the protein and does not form any hydrogen bonds with neighboring residues. However, four water molecules are within hydrogen bonding distance of the salt bridge, and waters appear in equivalent positions in all other T4 lysozyme structures in Protein Data Bank[31] (35 T4 lysozyme mutant structures from 1101.pdb to 1135.pdb by Matthews and co-workers obtained from the same crystal form). These water molecules can form hydrogen bonds with the ion pair when both Asp and His are charged (see Fig. 8).

Calculations were first carried out with the bound

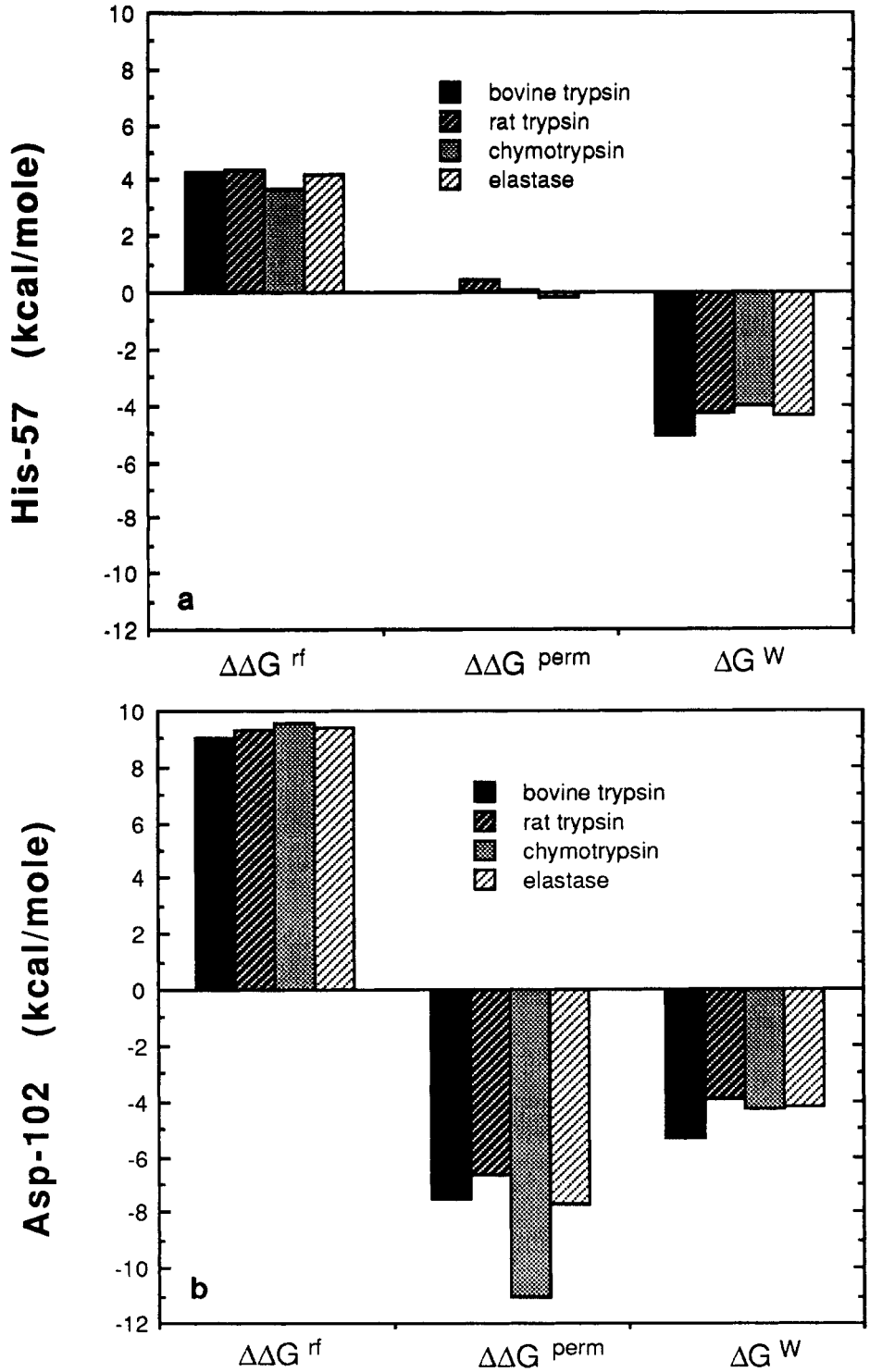

Fig. 7. (a) Calculated free energy terms, $\Delta\Delta G^{rf}$, $\Delta\Delta G^{perm}$, and ΔG^{W}, of His-57 for four serine proteases at pH 6 and (b) for Asp-102. Salt concentration 0.14 M.

waters treated as bulk waters, i.e., as a continuum with a dielectric constant of 80. The calculated pK_as were 3.3 for Asp-70 and 8.0 for His-31, the shifts for each being smaller than those observed experimentally. In order to explore the consequences of treating the bound waters as part of the bulk medium, we carried out a series of calculations where the bound waters were treated explicitly, i.e., as if they were

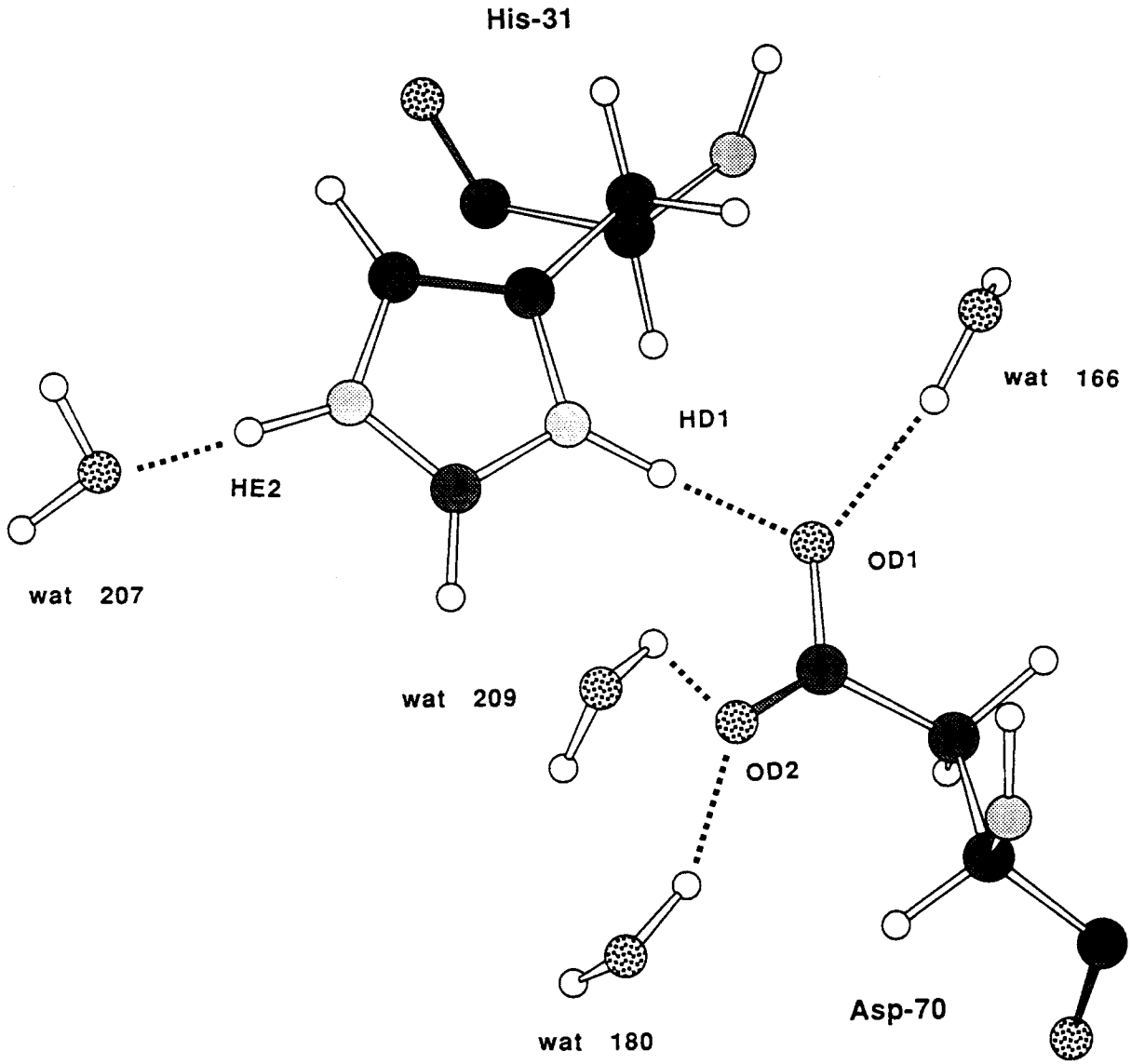

Fig. 8. The His-31 . . . Asp-70 salt bridge in T4 lysozyme with the crystal water molecules in hydrogen bonding positions.

part of the protein. An immediate complication arises with this procedure since the water positions are known only for the ionized form of the ion pair and there is no crystallographic information as to their location when one or both of these groups are neutral. Our approach has been to remove any explicit waters from the calculation that would no longer hydrogen bond to the neutral form of one of the residues. Thus for example, water 209 and 180 that hydrogen bond to OD2 of Asp-70 are turned into bulk water when OD2 picks up a proton in the unionized form of the residue.

Carrying out the calculations this way yields a pK_a near zero for Asp-70 and a pK_a of 9.9 for His-31. In this case, the calculated shifts are somewhat too large whereas treating the waters as bulk leads to shifts that are too small. These results may be understood from the following argument. Treating the waters explicitly maximizes the ion–water interaction, stabilizing the ionized form, because the energetic cost of binding these waters is not accounted for. In contrast, treating bound waters simply as a medium of dielectric 80 underestimates the stabilizing effects of a highly directed hydrogen bond as observed in T4 lysozyme.

DISCUSSION

We have reported a method to calculate pK_as in proteins based on FDPB electrostatic calculations and a new hybrid statistical mechanical/Tanford–

Roxby procedure to treat multiply interacting sites. The method is extremely fast, in part because the Tanford–Roxby iteration itself is found to work quite well in most cases. However, the Tanford–Roxby method does break down when ionizable groups are strongly coupled. In such cases, the statistical mechanical treatment of all interactions within a certain cutoff radius is found to correct the problems inherent in the use of Eq. (1) when more than one titratable site is involved. When compared to other methods, the hybrid method was found to be at least an order of magnitude faster than a Monte Carlo sampling technique, and orders of magnitude faster than the full or reduced statistical mechanical treatments when the number of ionizable sites is large (for example, apo-myoglobin).

Our results, as well as those reported previously, suggest that in most cases, it is possible to make reasonable predictions of pK_as in globular proteins. Our success in treating the His-57 . . . Asp-102 ion pair in serine proteases indicates that even complex situations such as buried salt bridges, which involve a sensitive balance between desolvation and compensating polar interactions, are amenable to theoretical analysis. In most cases encountered in globular proteins, the unfavorable desolvation penalty associated with removing an ionizable group from water is compensated for by favorable polar interactions with hydrogen bonding groups in the protein. In addition, most charge–charge interactions between ionizable groups on the surface of proteins are screened by the high dielectric solvent and by ionic strength effects. Both of these factors contribute to the general observation that the observed pK_as of most groups in proteins are quite close to those of the isolated amino acids.

An exception to this behavior noted in this work is the His-31 . . . Asp-70 salt-bridge in T4 lysozyme where both groups exhibit major pK_a shifts in opposite directions.[40] This behavior could not be accounted for without treating the four bound water molecules explicitly, which indicates a breakdown in the assumption of a dielectric continuum. This result highlights one of the uncertainties in all current treatments of pK_as in that it may be necessary to treat some water molecules as part of the protein rather than as part of the solvent. In some cases, high resolution crystal structures make it possible to identify these waters, but in cases where such information is not available, it would be necessary to predict their location. An additional uncertainty inherent in all pK_a calculations arises from the uncertainty of the three-dimensional structure of native proteins in solution, as shown by the difference between theory and experiment for the C-terminal of BPTI. Problems can arise from possible errors in the X-ray coordinates, conformational changes that may accompany protonation and deprotonation, uncertainty in the positions of polar hydrogens, and possible differences between crystal and solution conformations. Molecular dynamics or Monte Carlo simulations offer a partial solution to these complexities, as will be discussed in a future publication.

In summary, progress in the past few years has made it possible, within the framework of continuum electrostatics, to carry out pK_a calculations on essentially any system. Satisfactory results have been obtained for most cases studied so far although there have been a number of cases where the discrepancy between theory and experiment has been significant. In those instances, it is necessary to question the underlying assumptions in the model and to extract new information from its breakdown. One example discussed in this work is T4 lysozyme where the problem appeared attributable to the treatment of bound waters. In a recent study of thioredoxin, the failure to account for the anomalous pK_a of Asp-26 led to a testable hypothesis concerning the motion of another group.[41] These examples serve to highlight the current state of pK_a calculations. They appear to be reliable enough so that their failures are interesting.

ACKNOWLEDGMENTS

We thank Tim Standing and Robert Fletterick for providing rat trypsin coordinates, Anthony Nicholls for the continuing enhancement of DelPhi, and S. Sridharan for assistance with the Monte Carlo algorithm. This work was supported by NIH Grant GM-30518 and NSF Grant DMB-04384.

REFERENCES

1. Tanford, C. "Physical Chemistry of Macromolecules." New York, John Wiley, 1961.
2. Sharp, K. A., Honig, B. Electrostatic interactions in macromolecules: Theory and application. Annu. Rev. Biophys. Chem. 19:301–332, 1990.
3. Warshel, A., Aqvist, J. Electrostatic energy and macromolecular function. Annu. Rev. Biophys. Chem. 20:267–298, 1991.
4. Yang, A.-S., Honig, B. Electrostatic effects on protein stability. Curr. Opin. Struct. Biol. 2:40–45, 1992.
5. Tanford, C., Roxby, R. Interpretation of protein titration curves: Application to lysozyme. Biochemistry 11:2192–2198, 1972.
6. Tanford, C., Kirkwood, J. G. Theory of protein titration curves I. General equations for impenetrable spheres. J. Am. Chem. Soc. 79:5333–5339, 1957.
7. Matthew, J. B., Gurd, F. R. N., Garcia-Moreno, B., Flanagan, M. A., March, K. L., Shire, S. J. pH-dependent processes in proteins. CRC Crit. Rev. Biochem. 18:91–197, 1985.
8. Matthew, J. B., Gurd, F. R. Calculation of electrostatic interactions in proteins. Methods Enzymol. 130:413–453, 1986.
9. Russell, S. T., Warshel, A. Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. J. Mol. Biol. 185:389–404, 1985.
10. Warshel, A., Sussman, F., King, G. Free energy of charges in solvated proteins: Microscopic calculations using a reversible charging process. Biochemistry 25:8368–8372, 1986.
11. Gilson, M., Honig, B. Calculation of electrostatic potentials in an enzyme active site. Nature (London) 330:84–86, 1987.
12. Sternberg, M. J. E., Hays, F. R. F., Russell, A. J., Thomas,

P. G., Fersht, A. R. Prediction of electrostatic effects of engineering of protein charges. Nature (London) 330:86–88, 1987.
13. Bashford, D., Karplus, M. pK_as of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. Biochemistry 29:10219–10225, 1990.
14. Beroza, P., Fredkin, D. R., Okamura, M. Y., Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. Proc. Natl. Acad. Sci. U.S.A. 88:5804–5808, 1991.
15. Gilson, M., Honig, B. The dielectric constant of a folded protein. Biopolymers 25:2097–2119, 1986.
16. Bashford, D., Gerwert, K. Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. J. Mol. Biol. 224:473–486, 1992.
17. Bashford, D., Karplus, M. Multiple-site titration curves of proteins: An analysis of exact and approximate methods for their calculation. J. Phys. Chem. 95:9556–9561, 1991.
18. Gilson, M., Honig, B. The energetics of charge-charge interactions in proteins. Proteins 3:32–52, 1988.
19. Hagler, A. T., Huler, E., Lifson, S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. J. Am. Chem. Soc. 96:5319–5327, 1973.
20. Nicholls, A., Honig, B. A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J. Comp. Chem. 12:435–445, 1991.
21. Gilson, M., Sharp, K. A., Honig, B. Calculating the electrostatic potential of molecules in solution: Method and error assessment. J. Comp. Chem 9:327–335, 1987.
22. Marquart, M., Walter, J., Deisenhofer, J., Bode, W., Huber, R. The geometry of the reactive site and of the peptide groups in trypsin, typsinogen and complexes with inhibitors. Acta Crystallogr, Sect B 39:480, 1983.
23. Entry 4PTI, version of April 1987.
24. Entry 3PTB, version of March 1985.
25. Birktoft, J. J., Blow, D. M. The structure of crystalline alpha-chymotrypsin, the atomic structure of tosyl-alpha-chymotrypsin at 2 Å resolution. J. Mol. Biol. 68:187, 1972.
26. Entry 2CHA, version of May 1984.
27. Meyer, E., Cole, G., Radahakrishnan, R., Epp, O. Structure of native porcine pancreatic elastase at 1.65 angstrom resolution. Acta Crystallogr., Sect B 44:26, 1988.
28. Entry 3EST, version of July 1991.
29. Weaver, L. H., Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 angstrom resolution. J. Mol. Biol. 193:189, 1987.
30. Entry 2LZM, version of April 1988.
31. (a) Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. J Mol Biol. 112:535–542, 1977. (b) Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., Weng, J. Crystallographic Databases-Information Content, Software Systems, Scientific Applications. Bonn/Cambridge/Chester, Data Commission of the International Union of Crystallography, 1987.
32. Wuthrich, K., Wagner, G. Nuclear magnetic resonance of labile proteins in the basic pancreatic trypsin inhibitor. J. Mol. Biol. 130:1–18, 1979.
33. Wlodawer, A., Nachman, J., Gilliland, G. L., Gallagher, W., Woodward, C. Structure of form III crystals of bovine pancreatic trypsin inhibitor. J. Mol. Biol. 198:469–480, 1987.
34. Wlodawer, A., Deisenhofer, J., Huber, R. Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. J. Mol. Biol. 193:145–156, 1987.
35. Bell, J. A., Wilson, K. P., Zhang, X.-J., Faber, H. R., Nicholson, H., Matthews, B. W. Comparison of the crystal structure of bacteriophage T4 lysozyme at low, medium, and high ionic strength. Proteins 10:10–21, 1991.
36. Brown, L. R., DeMarco, A., Richarz, R., Wagner, G., Wuthrich, K. The influence of a single salt bridge on static and dynamic features of the globular solution conformation of the basic pancreatic trypsin inhibitor. Eur. J. Biochem. 88:87–95, 1978.
37. Warshel, A., Naray-Szabo, G., Sussman, F., Hwang, J. How do serine proteases really work? Biochemistry 28: 3629–3637, 1989.
38. Soman, K., Yang, A. S., Honig, B., Fletterick, R. Electrical potentials in trypsin isozymes. Biochemistry 28:9918–9926, 1989.
39. Fersht, A. "Enzyme Structure and Mechanism." New York: W. H. Freeman, 1985.
40. Anderson, D. E., Becktel, W. J., Dahlquist, F. W. pH-induced denaturation of proteins: A single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4-lysozyme. Biochemistry 29:2403–2408, 1990.
41. Langsetmo, K., Fuchs, J. A., Woodward, C., Sharp, K. A. Linkage of thioredoxin stability to titration of ionizable groups with perturbed pK_a. Biochemistry 30:7609–7614, 1991.
42. Richarz, R., Wuthrich, K. High-field 13C nuclear magnetic resonance studies at 90.5 MHz of the basic pancreatic trypsin inhibitor. Biochemistry 17:2263–2269, 1978.
43. Brown, B. R., DeMarco, A., Wagner, G., Wuthrich, K. A study of the lysyl residues in the basic pancreatic trypsin inhibitor using 1H nuclear magnetic resonance at 360 MHz. Eur. J. Biochem. 62:103–107, 1976.
44. DeMarco, A., Tschesche, H., Wagner, G., Wuthrich, K. Proton NMR studies at 360 MHz of the methyl groups in native and chemically modified basic pancreatic trypsin inhibitor. Biophys. Struct. Mech. 3:303–315, 1977.