Explicit Solvent Models in Protein pK_a Calculations

Cynthia J. Gibas and Shankar Subramaniam
Center for Biophysics and Computational Biology, Department of Molecular and Integrative Physiology, and Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA

ABSTRACT Continuum methods for calculation of protein electrostatics treat buried and ordered water molecules by one of two approximations; either the dielectric constant of regions containing ordered water molecules is equal to the bulk solvent dielectric constant, or it is equal to the protein dielectric constant though no fixed atoms are used to represent water molecules. A method for calculating the titration behavior of individual residues in proteins has been tested on models of hen egg white lysozyme containing various numbers of explicit water molecules. Water molecules were included based on hydrogen bonding, solvent accessibility, and/or proximity to titrating groups in the protein. Inclusion of water molecules significantly alters the calculated titration behavior of individual titrating sites, shifting calculated pK_a values by up to 0.5 pH unit. Our results suggest that approximately one water molecule within hydrogen-bonding distance of each charged group should be included in protein electrostatics calculations.

INTRODUCTION

The calculation of pK_a shifts of titratable residues in proteins is generally framed as a problem in two parts. If the pK_a of a model compound equivalent to the titrating residue i in the protein is known, then the shifted pK_a of titrating site i can be expressed as a function of the work of adding a charge q_i to the site:

$$\Delta G_i = \Delta G_{Born} + \Delta G_{back} + \sum_j q_i q_j W_{ij} \qquad (1)$$

where the three terms in the equation represent, respectively, the energy penalty of removing the residue from the bulk solvent, the difference between the interaction of the group i with the background of partial charges and dipoles in the protein and with the charges in residue i, and the interaction of the charge at i with charges at all other sites j. The last term, however, depends on the ionization states of the other charged groups with which i is interacting. The first problem, then, is the calculation of these energies. The complexity of protein systems requires that a numerical method be used for calculation of the energies, the most common method being to solve the Poisson-Boltzmann equation for the variation through space of the electrostatic potential due to a set of point charges embedded in a continuum electrolyte,

$$\nabla \cdot [\epsilon(r) \cdot \nabla\phi(r)] - \epsilon(r)\kappa^2(r)\phi(r) = -4\pi\rho(r) \qquad (2)$$

using a finite difference algorithm. This requires that the protein be modeled as an impermeable object of low dielectric constant in a medium of high dielectric constant, surrounded by an ion-inaccessible layer. The boundary between the two dielectric regions is taken to be either the Van der Waals surface or the solvent-accessible surface of the protein (Davis et al., 1991). Although this continuum model is a convenient one, it is not unambiguous. Well-resolved water molecules have been located in many x-ray crystal structures in proteins, whereas in most other structures, cavities are found into which one or many water molecules may be fitted.

The second problem is the calculation of the pK_a values of the individual titrating sites using these energies. This can be done by taking a Boltzmann-weighted sum over the possible protonation states or by making some approximation to this sum (Bashford and Karplus, 1990; Beroza et al., 1991; Oberoi and Allewell, 1993; Yang et al., 1993; Karshikoff, 1995); the ionization polynomial formalism, which yields equilibrium constants for each titrating site (Gilson, 1993; Antosiewicz et al., 1994), is equivalent to the above approach. Although several approximations to the problem of 2^N ionization states have been developed, the solutions are only as good as the calculated interaction energies. One element of protein structure that is ignored in continuum electrostatic models is the Coulombic interaction between atomic partial charges on the protein residues and those on water molecules trapped in cavities and clefts in the protein. The potential importance of these water molecules to electrostatics calculations has previously been discussed (Yang et al., 1993), although not investigated systematically.

It is conceptually incorrect to treat crystallographic water molecules, some of which have B-factors as low as those of main-chain atoms, as part of the high-dielectric continuum and to model water-containing cavities as having a dielectric constant of 80. Much of the high dielectric constant of water is due to orientational polarizability; in fields that oscillate at frequencies too high for water molecules to reorient with the field, water has a much lower apparent dielectric constant. Water molecules contained even in relatively large cavities in proteins are likely to have much less

Received for publication 16 February 1996 and in final form 15 April 1996.

Address reprint requests to Dr. Shankar Subramaniam, Beckman Institute University of Illinois at Urbana-Champaign, Urbana, IL 61801. Tel.: 217-244-4489; Fax: 217-244-2909; email: shankar@uiuc.edu.

0006-3495/96/07/138/10 $2.00

freedom of rotation than water molecules in the bulk solvent phase; thus, cavities in the protein should not merely be treated as part of the solvent dielectric region. Water molecules are found in predictable spatial relationships to polar and charged amino acids; their orientation and location are controlled by hydrogen bonding stereochemistry (Roe and Teeter, 1993; Williams et al., 1994). Thus, even water molecules at the protein-solvent interface may be considered to have more specific, local interactions with the protein than does the bulk solvent. In many proteins, water molecules are known to have important catalytic and structural roles. The classic example of the participation of bound water molecules in catalysis is that of the serine proteases. Antosiewicz et al. (1994) have used a continuum approach sans explicit water molecules to calculate pK_a values of active site residues in chymotrypsin (Antosiewicz et al., 1994). Despite their success, inclusion of a highly ordered chain of linked water molecules that stretches between Ser 214, a highly conserved residue near the catalytic triad, to the protein surface and bulk solvent may yield even more accurate pKa values. Catalytic activity can be blocked by mutations that obstruct this pathway (Meyer, 1992). Binding of water in the interior of bacteriorhodopsin is thought to cause conformational changes in the protein required for formation of the Schiff-base linkage between retinal and protein. Hydrogen bonding between water and protein may also enhance the nucleophilic attack by the amino group, which is required for Schiff-base formation (Rousso et al., 1995). The oxygen affinity of hemoglobin is known to vary linearly with the chemical potential of water in solution; binding of extra water molecules to the protein in a number sufficient to account for the extra solvent-accessible surface of the R state is a factor in the T (deoxy) to R (oxy) transition (Colombo et al., 1992). Even in the case of water molecules that are not functionally significant, it is intuitively incorrect to assume that the significant dipole of a fixed water molecule will have no effect on the electrostatic interactions of surrounding residues.

MATERIALS AND METHODS

Calculation of pK_a values

The method described in Gilson (1993) and Antosiewicz et al. (1994) was used to calculate pK_as of titrating residues in hen egg white lysozyme (HEWL), for which the pK_as of all titrating entities with the exception of Arg residues have been determined experimentally (Kuramitsu and Hamaguchi, 1980). This method implements the University of Houston Brownian Dynamics (Davis et al., 1991) suite of programs to calculate interaction energies. A modified Tanford-Roxby (1972) approach, in which an ionization polynomial is evaluated exactly for clusters of residues with significant charge-charge correlations, but in the mean field approximation for less significant intercluster interactions, is then used to determine pK_a values.

Tanford and Roxby (1972) first reported a computational approximation to the calculation of titration behavior, in which values for the average charge on each site are set initially based on the intrinsic pK_a values of each amino acid residue. For a given titrating site i interacting with other sites j in the protein, the pK_a shift to site i due to the influence of site j is

$$\Delta pK_{ij} = -\frac{W_{ij}}{2.303 z_i kT} \tag{3}$$

and the total shift in the pK_a value of the site is a summation of all of these shifts:

$$pK_i = pK_{int,i} + \sum_{j \neq i} pK_{ij} \tag{4}$$

The pK_a shifts calculated based on the initial charges are used to update the charges iteratively to convergence, and this is repeated at several pH values. This mean-field approximation has been shown to work well except in cases of strongly coupled sites titrating at similar pH values.

To increase the accuracy of the mean-field approximation, several methods in which local interactions are calculated exactly as statistical mechanical averages and long-range interactions are calculated using the mean-field approximation have been developed. Local interactions have been defined in two ways: either as residues falling within a specified cutoff radius (e.g., Yang and Honig, 1993) or as clusters of residues having large intercluster charge-charge correlations.

The method used here uses an interaction-energy clustering method and an ionization polynomial formalism. For a cluster I having nI titrating sites, the ionization polynomial is

$$\sum_I = \sum_{\alpha=0}^{2nI-1} \left(e^{-\beta G_{elec,\alpha}} e^{-\beta \sum_{i=1}^{nI} x_\alpha(i) \sum_{k \ni I} \theta_k G_{ik}} \prod_{i=1}^{nI} A_i^{x\alpha(i)} \right) \tag{5}$$

where the outer sum is over the ionization states of cluster I, and the $k \ni I$ indicates that k ranges over all clusters other than I, in summation of the mean influences of other clusters on the current cluster, as in the mean field approximation where the equilibrium of group i is influenced by group j as $\theta_j G_{ij}$. In this method, each group is initially set to its fully ionized state to avoid the failure of the mean field approximation to detect cooperative ionization of two initially neutral groups. Initial guess charges assigned by the Tanford-Roxby method are used for all residues outside of cluster I, and the ionization polynomial is evaluated and used to update the charges of groups in cluster I using the formula

$$\theta_i = \frac{1}{\sum_I} \sum_{\alpha=0}^{2nI-1} x_\alpha(i) e^{-\beta G_{elec,\alpha}} e^{-\beta \sum_{i=1}^{nI} x_\alpha(i) \sum_{k \ni I} \theta_k G_{ik}} \prod_{i=1}^{nI} A_i^{x_\alpha(i)} \tag{6}$$

for the fractional occupancies of all ionization states. The new charges are used in the computation of the fractional charges of the next cluster and so on, iteratively to self-consistency. Given a cluster size of 1 residue, this method is equivalent to the Tanford-Roxby approximation; generally a cluster size of 15 residues produces reasonable results (Gilson, 1993). Clusters are selected by a recursive search, in which clusters are chosen whose charge-charge interaction energies within the cluster are significant, but whose charge-charge interaction energies with residues in other clusters fall below a cutoff value. Initially the cutoff is set to zero, but it is incremented if clusters smaller than the cluster size limit cannot be found. Groups that are not likely to titrate at a particular pH are placed in single-residue clusters, allowing their treatment by mean-field methods on the assumption that inter-residue interactions will not significantly affect their ionization state at that pH.

Of the methods surveyed, This method produces the best agreement of calculated pK_a values with the pK_a values determined experimentally by Kuramitsu and Hamaguchi (1980), regardless of parameters used, as shown in Table 1. In the current application of the cluster method, charges were taken from the CHARMm (Brooks et al., 1982) parameter set and radii from the optimized parameters for liquid simulations (Jorgensen and Tirado-Rives, 1988) parameter set, following Antosiewicz et al. (1994). Calculations used a protein interior dielectric constant of 4, or of 20 (with which the original authors of the method obtained their best results). Four

TABLE 1 Calculated pK_a values for HEWL, compared to experimental values taken from (Ramanadham et al., 1981)

	expt.	B&K	B&F	O&A	Y&H	CLUS
LYSN_1	7.9	6.4	6.6	7.3	5.4	7.2
LYS_1	10.6	9.6	9.7	9.0	10.2	10.1
GLU_7	2.6	2.1	2.1	a.p.	3.4	2.7
LYS_13	10.3	11.6	11.7	a.p.	12.0	10.7
HIS_15	5.8	4.0	4.0	5.8	6.7	5.7
ASP_18	2.9	3.1	3.1	a.p.	4.0	2.8
TYR_20	10.3	14.0	o.r.	a.p.	n.r.	10.2
TYR_23	9.8	11.7	11.7	10.5	n.r.	9.2
LYS_33	10.4	9.6	9.6	11.2	10.2	11.0
GLU_35	6.1	6.3	6.3	5.7	2.6	4.6
ASP_48	4.3	1.0	1.0	6.2	1.6	3.5
ASP_52	3.55	7.0	7.0	4.3	5.2	2.5
TYR_53	12.1	20.8	o.r.	14.8	n.r.	11.1
ASP_66	2.0	1.7	1.7	2.3	3.1	1.2
ASP_87	3.65	1.2	1.2	3.6	1.6	2.4
LYS_96	10.7	10.4	10.4	10.2	10.5	11.9
LYS_97	10.1	10.6	10.6	a.p.	11.4	11.2
ASP_101	4.25	7.9	7.9	3.4	2.9	3.4
LYS_116	10.2	9.9	9.9	a.p.	10.6	10.1
ASP_119	2.5	3.2	3.3	a.p.	3.4	3.0
LEUC_129	2.9	2.3	2.3	a.p.	2.6	2.4
RMSD		2.6	1.8	1.1	1.7	0.7

B and K: Calculated values from Bashford and Karplus, 1990. B and F: Calculated values from Beroza et al., 1991. Values were not reported in tabular form and so were estimated from figure. o.r. indicates that the titration curve for that residue was out of range of the figure. Out of range values were not included in the root mean square deviation. O and A: Values from Oberoi and Allewell, 1993. Values were estimated from figures where possible, for comparison. a.p. indicates that the data point for that residue was ambiguous. RMSD is that reported by authors. Y and H: Values from Yang et al., 1993. Calculation did not include tyrosines so root mean square deviation is not strictly comparable. The cluster method calculation to which these are compared uses a protein dielectric constant of 4 and the standard continuum model.

focusing grids were used in the finite difference Poisson-Boltzmann calculations; the largest grid was a cubic grid of 102.5 Å with a grid spacing of 1.5 Å; the smaller grids of extents 18.0, 11.25, and 4.0 Å with grid spacings of 1.2, .75, and .25 Å, respectively, centered on the titrating site of interest. The solvent accessible surface was used as the boundary between the protein interior and exterior. A solvent ionic strength of 0.15 mM was used with radius 2.0 Å, and the temperature set at 298 K.

Selection of water molecules for inclusion in calculations

In this report, the effect of inclusion of various sets of explicit water molecules on the ionizable residues in proteins is shown for the small protein HEWL. The triclinic crystal structure of HEWL (2LZT) (Ramanadham et al., 1981) contains 249 water molecules, not all of which are equally likely to influence the titration behavior of the protein. Ten groups of water molecules selected from among these, using different restrictions for each set, are described in Table 2. Views of the structure of HEWL, which include these various subsets of water molecules, are shown in Fig. 1. In deciding whether a given water molecule was part of the continuum or part of a subset of water molecules that should be explicitly included in the atomic structure of the protein, several factors were considered. First, in the course of adding hydrogen atoms to the HEWL structure using the program HBUILD (Brunger and Karplus, 1988), which must be done before the calculation of pK_as, it was determined that 48 of the water molecules included in the structure were beyond the distance cutoff set in HBUILD for an acceptor search to be carried out. The water molecules were too distant from the protein surface to be oriented by it using this method. These water molecules were removed from the structure; the remaining 201 water molecules comprise explicit solvent model A.

TABLE 2 Subsets of water molecules included in the input PDB files for these calculations, characterized by the number of water molecules included and the selection criteria used to choose each subset

water molecule subset	# water molecules	method of selection
A	201	all water molecules within cutoff distance for h–bond acceptor search in HBUILD
B	105	all water molecules with hydrogen bonds to protein
Q	103	all water molecules within 5A of a titrating site
C	33	all water molecules <20% exposed to bulk solvent
D	15	all water molecules <10% exposed to bulk solvent
E	17	all water molecules <20% exposed to bulk solvent and within 5A of a titrating site
F	6	all water molecules <10% exposed to bulk solvent and within 5A of a titrating site
G	6	all water molecules 0% exposed to bulk solvent
H	1	all water molecules 0% exposed to bulk solvent and within 5A of a titrating site
N	0	

Each set is assigned an arbitrary letter code, which is consistently used throughout this report to refer to the set in question.

The fractional solvent accessibility of each water molecule was also considered. This was determined using the program VOLBL (Liang et al., in preparation). When water molecules are included in the protein structure used in an accessibility calculation, some water molecules may be occluded from the solvent by others and thus appear to be buried. To select a subset of water molecules with fractional accessibility less than a specified cutoff value, a serial procedure was used in which accessibility was calculated, followed by the removal of all water molecules with fractional accessibility greater than the cutoff value, iteratively until no further water molecules could be removed. Explicit solvent models C, D, and G are subsets of set A selected on the basis of solvent accessibility (see Table 2) using 20%, 10%, and zero accessibility as cutoffs, respectively.

Hydrogen-bonding interactions and proximity to titrating sites were also used to restrict the number of water molecules included in the calculations. Hydrogen bonding interactions were calculated using a donor-acceptor distance cutoff of 3.3 Å; and water molecules in proximity to titrating sites (found within a 5-Å radius of the titrating site) were selected using Quanta 4.1 (Molecular Simulations, Inc., Waltham, MA). Explicit solvent model B is a subset of set A selected on the basis of hydrogen bonding to the protein, whereas explicit solvent model Q is a subset of set A selected on the basis of proximity to titrating sites. Though models B and Q contain similar numbers of water molecules, they are substantially different; water molecules in model B are quite evenly distributed, whereas those in model Q are dense over some parts of the HEWL molecule and sparse over other parts. This can be seen in Fig. 1. The proximity criterion was also used to further limit the number of water molecules found in solvent models C, D, and G, yielding solvent models E, F, and H.

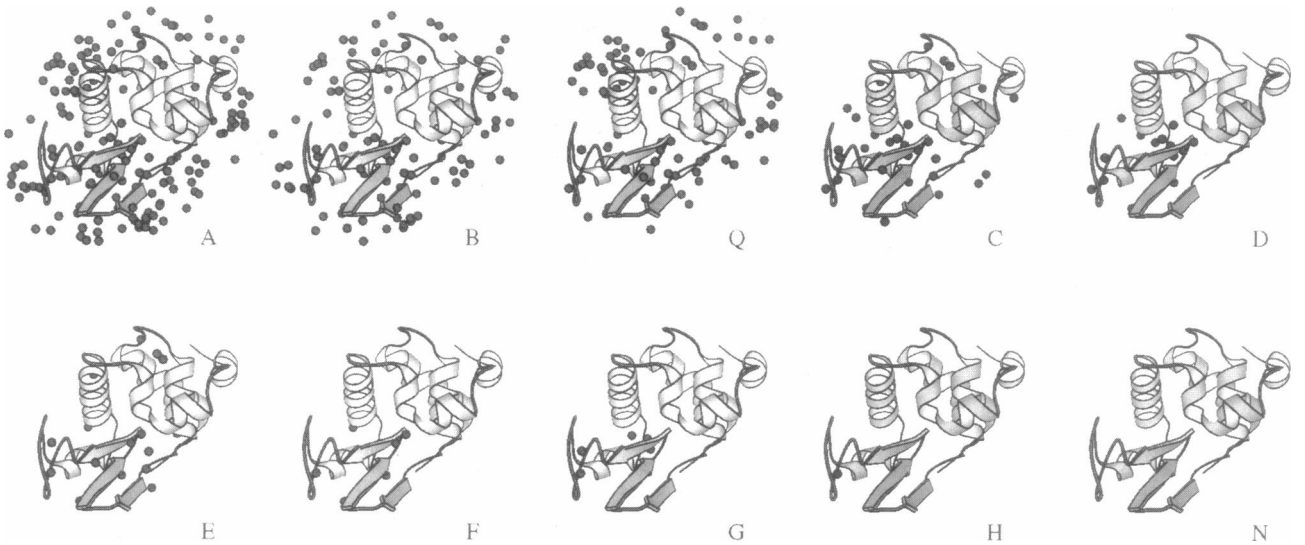

FIGURE 1 Ten views of HEWL in the same orientation, with the various groups of water molecules used in these calculations shown. Water models A-N are as described in Table 2. Figures were generated using Molscript (Kraulis, 1991)

Crystallographic B-factors are useful in determining how well-ordered an atom in a protein structure can be considered to be. B-factors of main-chain atoms in lysozyme range from 9 in well-ordered regions of the structure to 30 in loop regions and at the N and C termini. In all but the three largest sets of water molecules described in Table 2, no water molecules are included that have crystallographic B-factors greater than 35; in the smaller sets the vast majority of included water molecules have B-factors less than 20.

Calculation of solvent accessibility of titrating sites

The molecular surface of each water-containing structure was calculated using the α shape software developed by Edelsbrunner et al. (1995). Differences in the fractional accessibilities of individual residues between the water-containing structures and the default, nonwater-containing structure were computed. As the coordinates of the atoms of the protein

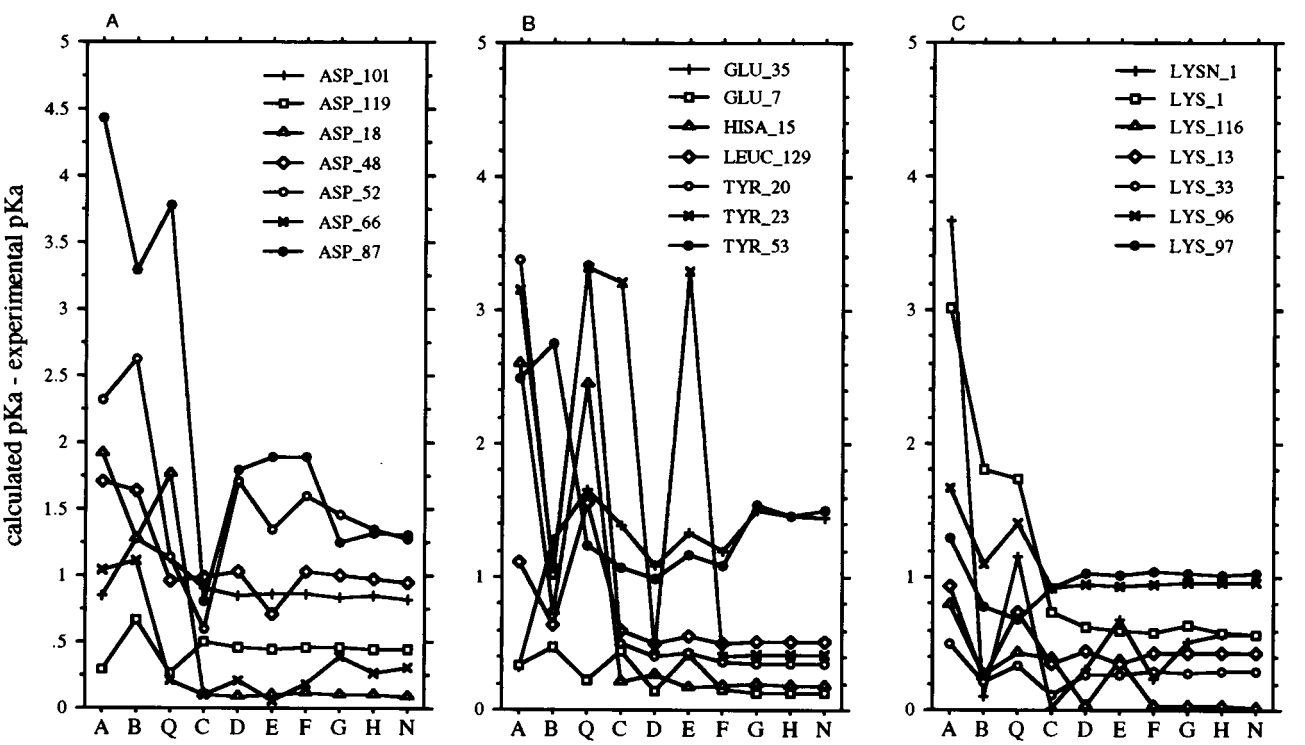

FIGURE 2 Absolute difference between calculated pK_a values and experimental values versus solvation model, for the 21 titrating sites in HEWL for which comparison to experiment is possible. Plots for the specific titrating sites have been distributed among three figures for clarity. *A*) Asp residues, *B*) Glu, His, Tyr, and Lys residues and C-terminus, *C*) Lys residues and N terminus. Water models A-N are as described in Table 2.

TABLE 3 Calculated pK_a values for the 21 experimentally characterized titrating sites of lysozyme under a variety of solvation conditions are compared to those calculated using the standard continuum model

		water molecule subset									
site	expt. pKa	A	B	Q	C	D	E	F	G	H	N
N-term	7.9	4.2	7.8	6.7	7.9	7.6	8.6	7.7	7.4	7.3	7.3
Lys_1	10.6	7.6	8.8	8.9	9.9	10.0	10.0	10.0	10.0	10.0	10.0
Arg_5		11.3	11.5	11.2	12.5	12.6	12.6	12.6	12.6	12.6	12.6
Glu_7	2.6	2.3	2.1	2.4	3.0	2.7	3.0	2.7	2.7	2.7	2.7
Lys_13	10.3	9.4	10.1	9.6	10.6	10.7	10.6	10.7	10.7	10.7	10.7
Arg_14		12.5	12.4	12.4	12.2	12.2	12.3	12.2	12.2	12.2	12.2
His_15	5.8	3.2	5.1	3.3	5.6	5.5	5.6	5.6	5.6	5.6	5.6
Asp_18	2.9	1.0	1.6	1.1	2.8	2.8	2.8	2.8	2.8	2.8	2.8
Tyr_20	10.3	13.7	9.3	13.6	9.8	9.9	9.9	10.0	10.0	10.0	10.0
Arg_21		12.6	13.1	12.6	13.1	13.2	13.1	13.2	13.2	13.2	13.2
Tyr_23	9.8	13.0	10.9	13.1	13.0	9.3	13.1	9.4	9.4	9.4	9.4
Lys_33	10.4	9.9	10.2	10.1	10.5	10.7	10.7	10.7	10.7	10.7	10.7
Glu_35	6.1	5.8	4.8	4.5	4.7	5.0	4.8	4.9	4.6	4.6	4.7
Arg_45		12.1	11.6	11.8	12.1	12.1	12.0	12.1	12.1	12.2	12.1
Asp_48	4.3	2.6	2.7	3.3	3.3	3.3	3.6	3.3	3.3	3.3	3.4
Asp_52	3.6	1.2	0.9	2.4	3.0	1.8	2.2	2.0	2.1	2.2	2.3
Tyr_53	12.1	9.6	9.4	10.9	11.0	11.1	10.9	11.0	10.5	10.6	10.6
Arg_61		13.1	13.2	13.3	13.5	13.1	13.6	13.2	13.1	13.2	13.2
Asp_66	2.0	1.0	0.9	1.8	2.1	2.2	2.1	2.2	1.6	1.7	1.7
Arg_68		15.1	14.6	15.4	14.8	14.7	14.7	14.7	14.4	14.4	14.4
Arg_73		12.4	12.4	12.3	12.3	12.2	12.2	12.2	12.2	12.2	12.2
Asp_87	3.7	−0.8	0.4	−0.1	2.9	1.9	1.8	1.8	2.4	2.3	2.3
Lys_96	10.7	9.0	11.8	9.3	11.6	11.7	11.6	11.6	11.7	11.7	11.7
Lys_97	10.1	11.4	10.9	10.8	11.0	11.1	11.1	11.1	11.1	11.1	11.1
Asp_101	4.3	3.4	3.0	3.1	3.3	3.4	3.4	3.4	3.4	3.4	3.4
Arg_112		12.2	12.3	12.2	12.4	12.5	12.5	12.5	12.6	12.5	12.5
Arg_114		12.8	12.5	12.1	12.3	12.4	12.3	12.4	12.4	12.4	12.4
Lys_116	10.2	9.4	9.9	9.8	9.8	10.2	9.9	10.2	10.2	10.2	10.2
Asp_119	2.5	2.8	3.2	2.8	3.0	3.0	2.9	3.0	3.0	2.9	3.0
Arg_125		12.6	12.7	12.5	12.7	12.7	12.7	12.7	12.7	12.7	12.7
Arg_128		12.2	12.2	12.2	12.2	12.2	12.1	12.1	12.1	12.1	12.1
C-term	2.9	1.8	2.3	1.3	2.3	2.4	2.3	2.4	2.4	2.4	2.4
rmsd		2.0	1.2	1.6	1.0	.8	1.1	.8	.8	.7	.7
mean ΔpKa		1.8	1.2	1.4	.7	.7	.8	.7	.7	.7	.7

Water models A–N are as described in Table 2.

molecule are the same in each of these structures, the calculated differences in surface area translate to the fraction of surface area occluded by explicit water molecules in the water-containing structures. α shape theory provides a method for analytically computing the molecular surface areas of proteins. In this method, a simplicial complex called an α-complex is derived from the Delaunay triangulation of the set of points that make up a protein structure. The Delaunay triangulation is based upon the Voronoi decomposition of the space filled by the protein molecule, in which the space is divided into cells containing one atom each, with each atom

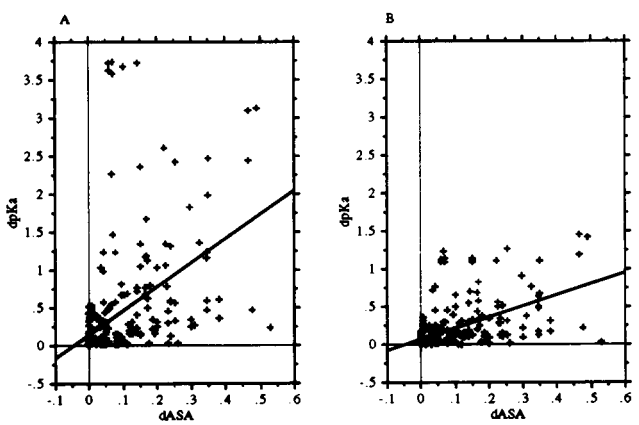

FIGURE 3 Absolute deviation of calculated pK_a values from those calculated for the default model (N, 0 waters) plotted as a function of the absolute difference in solvent accessibility. Results from all nine water-containing models are included. When all datasets are considered, Spearman's rho is .755 and Z is 12.79.

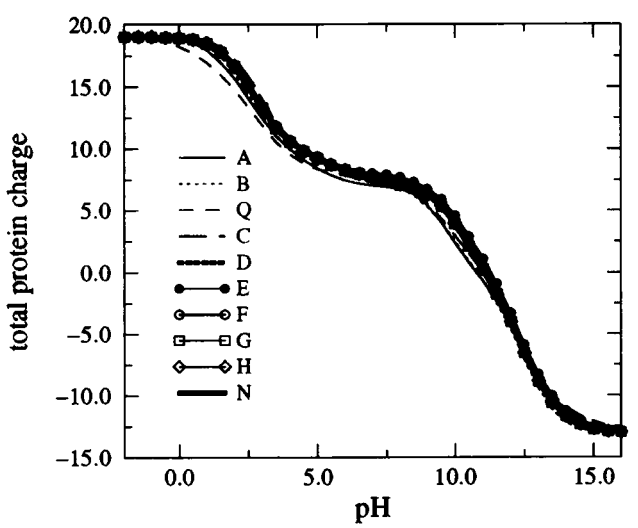

FIGURE 4 Calculated whole protein titration curves for HEWL with various explicit solvent models as described in Table 2.

TABLE 4 A summary of local interactions of water molecules with individual titrating sites in HEWL as they vary with the explicit solvent model chosen

	water molecule neighbors in model					
site	A,Q only	A,Q, and B	A,Q,B,C,E	A–F	all models	best calculated pKa value, model
N–term	149, 216	149, 216	216			C
Lys_1	143,241	143,241				D–N
Glu_7	144,145,166	144,145,166	166			D, F–N
Lys_13	165,226	165,226	165			B
His_15	317					C, E–N
Asp_18	165,184,253	165,253	165			C–N
Tyr_20	174,179	179				F–N
Tyr_23	183	183	183			F–N
Lys_33	233,282	233,282				C
Glu_35	172,246	172	172			A
Asp_48	150,261	150	150			E
Asp_52	206,345,372	206,345,372	206			C
Tyr_53	135,151	135,151	135			D
Asp_66	137,139,187	137,139,187	137,139	137,139	139	C, E
Asp_87	142,158,180 207,317,323	142,158,180 207,323	142,180	180		C
Lys_96	237,302	302				B
Lys_97	286,315,363 366,368	286,315,363 368				Q
Asp_101	163,368	163,368				A, D–N
Lys_116	287					D, F–N
Asp_119	190,276	190,276				A, Q
C–term						D, F–N

The first column, labeled "A,Q only," gives all the water neighbors of each titrating site, indicated by their number in the original Protein Data Bank structure 2LZT. Within hydrogen bonding distance of the titrating sites, set A and set Q are equivalent. The second column gives only those water molecules that are also present in model B, the third column those that are also present in models C and E (which are equivalent within hydrogen bonding distance of the titrating sites), and so on. The final column gives the explicit solvent model that produced the best calculated pk_a value for that site. Listing of more than one model in this column indicates that results for several of the models were very similar.

"owning" all the space that is closer to that atom than to any other atom. Atoms in the Voronoi diagram correspond to vertices in the Delaunay triangulation: two cells that share a face to a line connecting the vertices, three cells having a common intersection to a triangle connecting the three vertices, etc. The simplicial complex is a complicated object built up from the points, lines, triangles, and tetrahedra of the Delaunay triangulation, and it is combinatorially equivalent to the actual molecule when the radii assigned to each atom center in the Voronoi decomposition are equal to the actual Van Der Waals radii. Once the α-complex is constructed, algorithms for inclusion and exclusion can be used to exactly determine any metric property of the protein structure. (Liang et al., in preparation)

The fractional solvent accessibility of an amino acid residue is defined as the measured solvent-accessible surface area of the residue in the protein divided by a standard state surface area. In the interest of consistency, standard state areas for each residue type were calculated using VOLBL for use in determining the fractional accessibilities discussed herein. The extended standard state area of each residue was defined as the area of the central residue in a Gly-X-Gly tripeptide with dihedral angles F = -140°, Ψ = 135°, C^{1} = -120°, and $C^{2\text{-}N}$ = 180°, as described in (Lesser and Rose, 1990).

RESULTS AND DISCUSSION

Results are shown graphically in Fig. 2 for those calculations in which a protein dielectric of 4 was used. Addition of water molecules in excess of those inaccessible to the bulk phase proved unnecessary and in fact detrimental to the overall accuracy of the calculations. Inclusion of subsets made up of completely or mainly solvent-inaccessible water molecules resulted in root mean squared deviations from the experimental values similar to those obtained when no water molecules were used, while improving specific results for several residues. Calculated pK_a values are given in Table 3. Addition of water molecules generally resulted in very slight changes in the pK_a values calculated for many residues; several values were substantially closer to experimental values, i.e., by 0.5 pH unit, than when no explicit water molecules were used, but one or two values were also shifted substantially away from the experimental values when explicit water molecules were used.

The change in calculated per-residue fractional solvent accessibility when explicit water molecules are added to a Protein Data Bank file is a measure of how much of the residue's surface is occluded by explicit solvent. Explicit solvent is considered to be part of the low-dielectric region; their inclusion moves the dielectric boundary (taken to be the solvent-accessible surface in UHBD calculations) away from the atoms that they occlude. Thus the change in fractional accessibility can be taken as a rough measure of the change in the individual residue's dielectric environment when explicit solvent is added to the model system. Figs. 3, *A* and *B* show absolute change in calculated pK_a as a function of change in fractional solvent accessibility between structures with water molecules included and the structure with no water molecules. Results from comparison of all nine water models with the null model are included in the plots; Spearman's ρ for this dataset is .755 regardless of

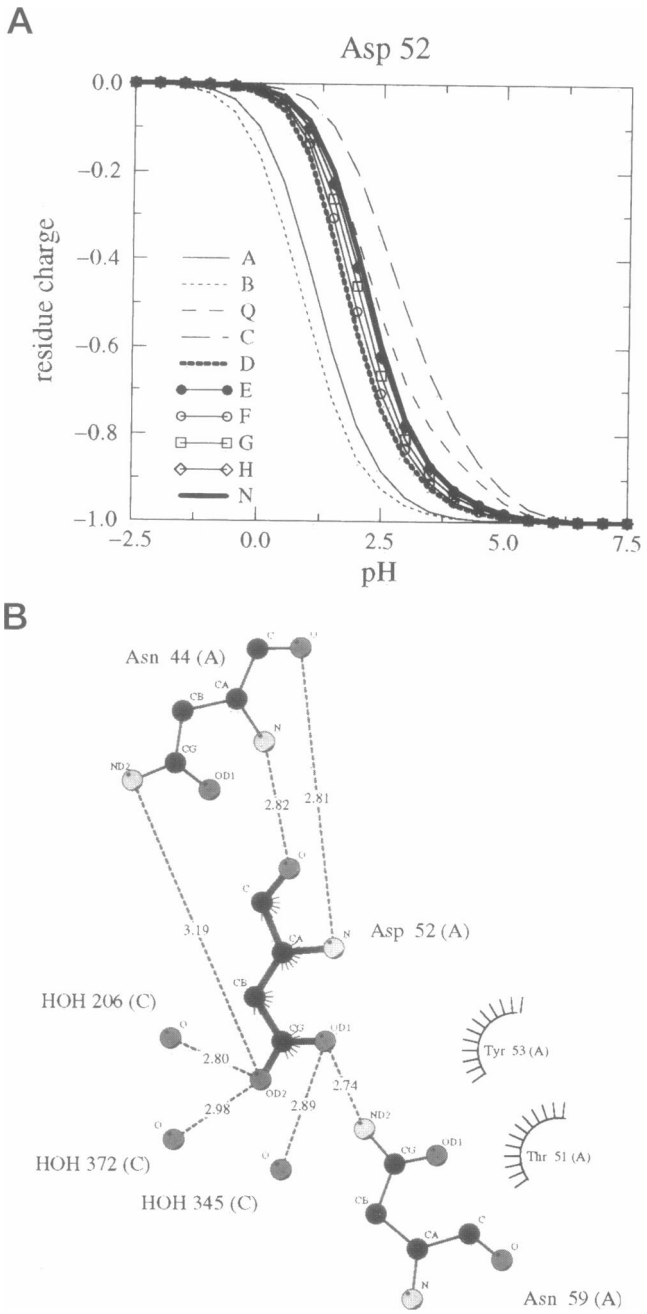

FIGURE 5 *A*) Titration curve for Asp 52. *B*) Local environment of Asp 52, including all water molecules within hydrogen-bonding distance. Not all of these water molecules are included in all of the explicit solvent models, as discussed in the text. Best calculated results are obtained for this residue when only HOH 206 is included in its immediate environment.

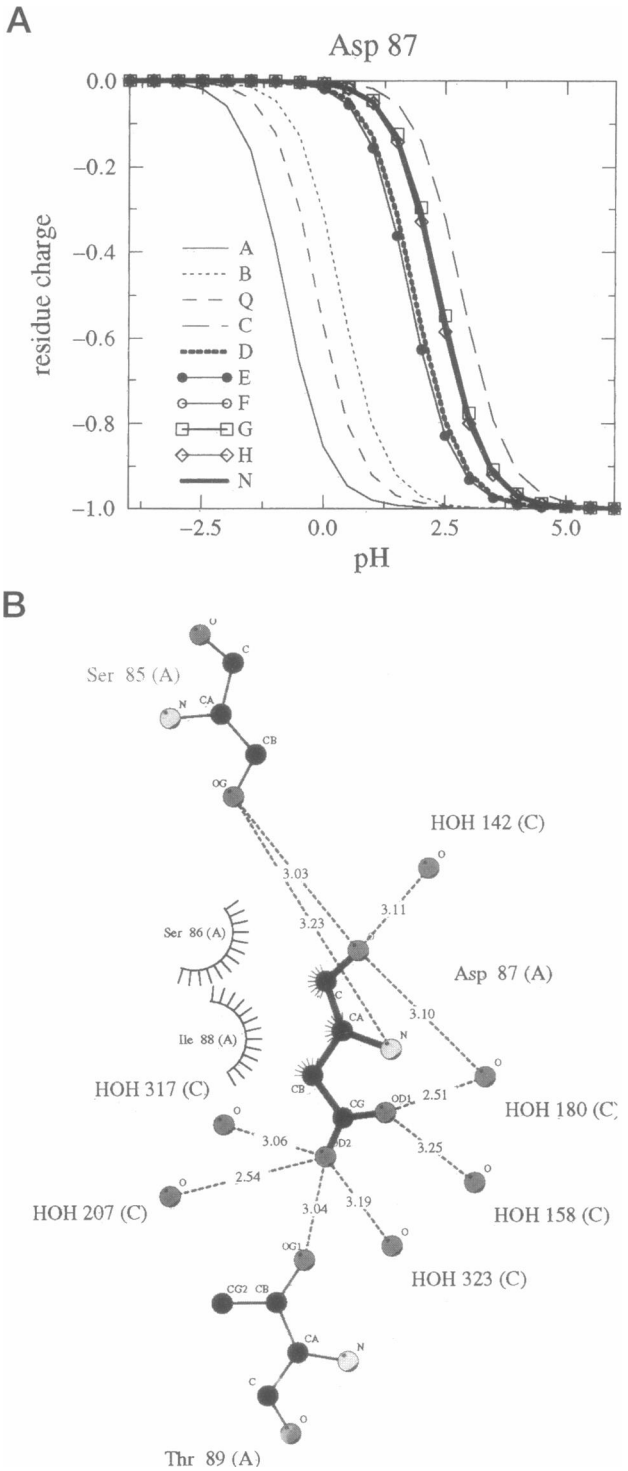

FIGURE 6 *A*) Titration curve for Asp 87. *B*) Local environment of Asp 87, including all water molecules within hydrogen-bonding distance. Best calculated results are obtained for this residue when only HOH 180 is included in its immediate environment.

whether an internal dielectric of 4 or of 20 is used in the calculation; the Z-value is 12.79 when an internal dielectric of 4 is used and 13.13 when an internal dielectric of 20 is used. This correlation indicates that including a particular water molecule in the pK_a calculation most significantly affects the titration behavior of residues with which the water molecule is in immediate contact. Thus individual water molecules in a region of particular interest in a protein structure can be included in pK_a calculations without perturbing the overall titration behavior of the protein significantly. Calculated whole-protein titration curves for HEWL with various explicit-solvent models are shown in Fig. 4.

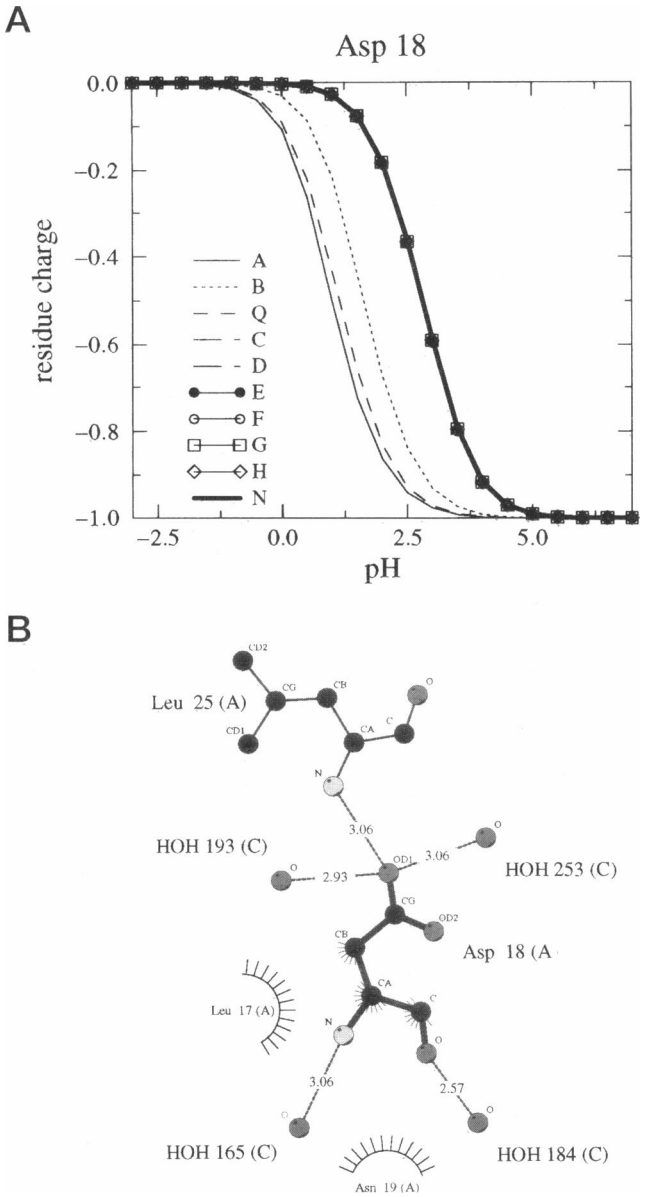

FIGURE 7 *A*) Titration curve for Asp 18. *B*) Local environment of Asp 18, including all water molecules within hydrogen-bonding distance. Best calculated results are obtained for this residue when no water molecules are included in its immediate vicinity, though the effects of removing only one or the other of the water molecules near atom OD1 has not been investigated.

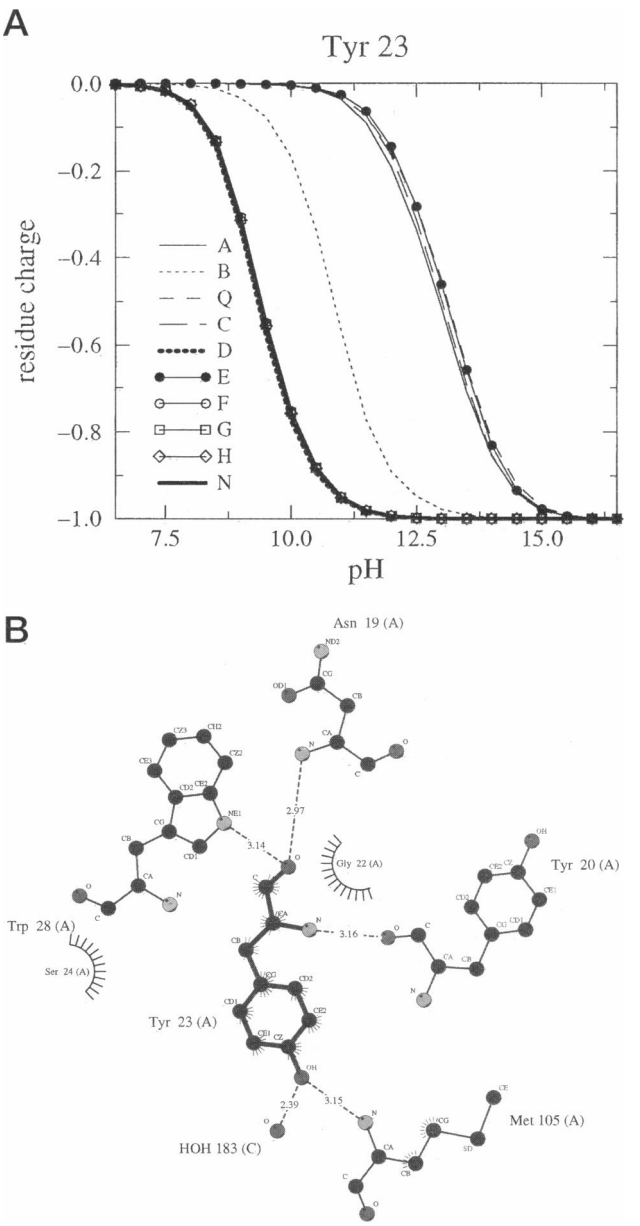

FIGURE 8 *A*) Titration curve for Tyr 23. *B*) Local environment of Tyr 23, including all water molecules within hydrogen-bonding distance. Best calculated results are obtained for this residue when no nearby water molecules are included.

The calculated titration behavior of the protein is very similar regardless of the explicit solvent model used; in the cases of models F, G, and H, which contain very few explicit water molecules, they are virtually identical. The most significant alteration to the protein titration curve occurs when solvent model Q is used. Use of this model, in which all water molecules to be found within 5 Å of titrating sites are included, results in a broadening and flattening of the protein titration curve.

In terms of prediction of individual pK_a values, none of the explicit solvent models considered was definitively the most correct. Model C, which contained only water molecules that were <20% buried, had a somewhat larger root mean square deviation than the models with smaller numbers of water molecules. However, this model produced the largest number of best calculated pK_a values for individual sites, followed closely by the model in which no water molecules were included (N). As there is very little difference in the quality of overall results unless a large number of explicit water molecules are included (i.e., in models A, B, and Q), the immediate environments of individual titrating sites and their titration behavior in the different explicit solvent models were examined to determine what kind of

relationship between individual titrating sites and water molecules might indicate that a water molecule could usefully be included in electrostatics calculations.

A summary of local interactions with water molecules for each titrating site is given in Table 4. Calculated titration curves and local hydrogen bonding environments for several of the residues discussed below are shown in Figs. 5-8. Residue environments were determined using the program HBPLUS (McDonald and Thornton, 1994) for calculation of hydrogen bonds, and the related plotting package LIGPLOT (Wallace et al., 1995), which reduces a 3-dimensional structure of a residue's environment to a 2-dimensional schematic.

Several titrating residues, including the N-terminus, Lys 13, Glu 35, Asp 48, Asp 52, Tyr 53, Asp 66, Asp 87, and Asp 119, were characterized by an environment having several water molecules as potential hydrogen bonding partners. In general, only one of these water molecules was potentially hydrogen bonded to the titrating group in the residue, or if more than one water molecule was within hydrogen bond distance of the titrating group, then only one of the water molecules was <20% buried. For this type of group, calculated pK_a values agreed best with experimental values in the case (usually explicit solvent model C, though in some cases explicit solvent model Q) in which only the buried water molecule in proximity to the titrating group was used. Agreement between calculated and experimental pK_a values worsened when this water molecule was not included. This type of site is exemplified in Figs. 5 and 6.

A second type of site, including His 15, Tyr 23, Lys 33, Lys 116, and the C-terminus, interacts with only one solvent molecule near their titrating group (0 in the case of the C-terminus), and the best calculated pK_a values for these residues are achieved in the absence of solvent. This type of site is exemplified in Fig. 7

Another type of site, including Lys 1, Glu 7, Asp 18, Tyr 20, Lys 97, and Asp 101 were also characterized by environments containing more than one water molecule hydrogen bonded to the titrating group of the residue. However, in these cases, the water molecules were not distinguishable from each other on the basis of solvent accessibility. They were generally present only in models A, B, and Q, and not present in any of the smaller sets, making it impossible to determine from the data which if any of the water molecules should be included. Generally, pK_as for these sites were found to be overshifted when solvent models A, B, and Q are used and undershifted when no water molecules are included in their immediate environment, with the usual result that the best calculated pK_a value for that site was found either when no water molecules were used or when one of the largely indistinguishable small solvent sets was used. It would be instructive to create more solvent models in which only one water molecule is included near each of these sites. This type of site is exemplified in Fig. 8.

It is our conclusion that the inclusion of explicit solvent molecules in the protein models used for electrostatics calculation improves the accuracy of the models. Although the inclusion of a select set of water molecules does not necessarily improve the overall root mean square deviation of the pK_a shifts, individual titratable residues with ordered hydrogen-bonded water molecules show significantly better correlation with experimental pK_a values when these water molecules are explicitly included in the calculation. As an additional caveat, we wish to point out that pK_a values could depend on orientation of the water molecules in the different states. The ideal number of solvent molecules is suggested to be something more than the smallest models used here and something less than the largest. Selection of a single water molecule in close proximity with each titrating site, whether the water molecule is solvent-accessible or not, may give the most correct representation of water effects on the behavior of individual titrating sites.

The authors thank Dr. Andrew McCammon for providing us the University of Houston Brownian Dynamics suite of programs, and the anonymous reviewers for their helpful comments. This work was supported by grants from the National Science Foundation (S.S.), a metacenter computer allocation (S.S.), and Graduate Assistance in Areas of National Need graduate research fellowship (C.J.G.)

REFERENCES

Antosiewicz, J., J. A. McCammon, and M. K. Gilson. 1994. Prediction of pH-dependent properties of proteins. *J. Mol. Biol.* 238:415–436.

Bashford, D., and M. Karplus. 1990. Multiple-site titration curves of proteins: an analysis of exact and approximate methods for their calculation. *Biochemistry.* 29:10219–10225.

Beroza, P., D. R. Fredkin, M. Y. Okamura, and G. Feher. 1991. Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of *Rb. Sphaeroides. Proc. Nat. Acad. Sci.* 88:5804–5808.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1982. CHARMm: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

Brunger, A. T., and M. Karplus. 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins Struct. Funct. Genet.* 4:148–156.

Colombo, M. F., D. C. Rau, and V. A. Parsegian. 1992. Protein solvation in allosteric regulation: a water effect on hemoglobin. *Science.* 256: 655–659.

Davis, M. E., J. D. Madura, B. A. Luty, and J. A. McCammon. 1991. Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program. *Comput. Phys. Commun.* 62:187–197.

Edelsbrunner, H., M. Facello, P. Fu, and J. Liang. 1995. Measuring proteins and voids in proteins. *In* Proceedings of the 28th Annual Hawaii International Conference on System Sciences. 256–264.

Gilson, M. K. 1993. Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins. *Proteins Struct. Funct. Genet.* 15:266–282.

Jorgensen, W. L., and J. Tirado-Rives. 1988. The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Amer. Chem. Soc.* 110:1657–1666.

Karshikoff, A. 1995. A simple algorithm for the calculation of multiple site titration curves. *Prot. Eng.* 8:243–248.

Kraulis, P. J. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24:946–950.

Kuramitsu, S., and K. J. Hamaguchi. 1980. Analysis of the acid-base titration curve of hen lysozyme. *Biochemistry.* 87:1215–1219.

Lesser, G. J., and G. D. Rose. 1990. Hydrophobicity of amino acid subgroups in proteins. *Proteins: Struct. Funct. Genet.* 8:6–13.

McDonald, I. K., and J. M. Thornton. 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238:777–793.

Meyer, E. 1992. Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. *Prot. Sci.* 1:1543–1562.

Oberoi, H., and N. M. Allewell. 1993. Multigrid solution of the nonlinear Poisson-Boltzmann equation and calculation of titration curves. *Biophys. J.* 65:48–55.

Ramanadham, M., L. C. Sieker, and L. H. Jensen. 1981. Structure of triclinic lysozyme and its $Cu(2^+)$ complex at 2 Å resolution. *Acta Crystallogr.* 37C:33.

Roe, S. R., and M. M. Teeter. 1993. Patterns for prediction of hydration around polar residues in proteins. *J. Mol. Biol.* 229:419–427.

Rousso, I., I. Brodsky, A. Lewis, and M. Sheves. 1995. The role of water in retinal complexation to bacterio-opsin. *J. Biol. Chem.* 270: 13860–13868.

Tanford, C., and R. Roxby. 1972. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry.* 11:2192–2198.

Wallace, A. C., R. A. Laskowski, and J. M. Thornton. 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Prot. Eng.* 8:127–134.

Williams, M. A., J. M. Goodfellow, and J. M. Thornton. 1993. Buried waters and internal cavities in monomeric proteins. *Prot. Sci.* 3:1224–1235.

Yang, A., M. R. Gunner, R. Sampogna, and B. Honig. 1993. On the calculation of pK_as in proteins. *Proteins Struct. Funct. Genet.* 15: 252–265.