Very Fast Empirical Prediction and Rationalization of Protein pK_a Values

Hui Li,[1†] Andrew D. Robertson,[2] and Jan H. Jensen[1*]
[1]*Department of Chemistry and Center for Biocatalysis and Bioprocessing, The University of Iowa, Iowa City, Iowa*
[2]*Department of Biochemistry, The University of Iowa, Iowa City, Iowa*

ABSTRACT **A very fast empirical method is presented for structure-based protein pK_a prediction and rationalization. The desolvation effects and intra-protein interactions, which cause variations in pK_a values of protein ionizable groups, are empirically related to the positions and chemical nature of the groups proximate to the pK_a sites. A computer program is written to automatically predict pK_a values based on these empirical relationships within a couple of seconds. Unusual pK_a values at buried active sites, which are among the most interesting protein pK_a values, are predicted very well with the empirical method. A test on 233 carboxyl, 12 cysteine, 45 histidine, and 24 lysine pK_a values in various proteins shows a root-mean-square deviation (RMSD) of 0.89 from experimental values. Removal of the 29 pK_a values that are upper or lower limits results in an RMSD = 0.79 for the remaining 285 pK_a values. Proteins 2005;61:704–721.**

INTRODUCTION

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

Theoretical interpretation and prediction of protein pK_a values are useful for understanding many biochemical problems. Ionizable groups with unusually low or high pK_a values tend to occur at protein active sites,[2–4] so identification of unusual pK_a values may facilitate the identification of protein/enzyme active sites,[5,6] as well as their functional mechanisms.[3,7] Rational design of drugs and new proteins will benefit tremendously from fast pK_a prediction methods. Some theoretical studies of proteins, such as force field simulations, need the pK_a values of ionizable groups to determine their charge states.

The most popular pK_a prediction methods for proteins[8–13] are based on electrostatic continuum models that numerically solve the linearized Poisson-Boltzmann equation (LPBE). In these methods, the protein is treated by a molecular mechanics force field, embedded in a uniform dielectric continuum with dielectric constants of 80 for the solvent and 4–20 for the protein interior. A pK_a shift is calculated from the difference in electrostatic energy of a residue in its charged and neutral form and this shift is added to a model pK_a value. These models[8,14–17] usually have a root-mean-square deviation (RMSD) from experiment of <1. However, the currently available LPBE models tend to overestimate the intraprotein charge–charge interactions[18,19] and underestimate the hydrogen-bonding and desolvation effects in calculating pK_a shifts.[20] The LPBE methods typically require tens of minutes to hours of computer time.

The pK_a values of small organic acids and bases can be predicted with the Hammet-Taft method[21] based on empirical relationship between pK_a shifts and substituents. The Hammet-Taft method is accurate for molecules similar to the ones included in the parameterization and is very fast. However, it is not generally applicable to proteins because the protein pK_a shifts are not solely due to substituent effects.

In this article, we present a study on the empirical relationships between protein pK_a shifts and structures. These empirical relationships are in turn used for structure-based protein pK_a prediction and rationalization. We show that this empirical method is able to accurately predict most protein pK_a values in a matter of seconds.

This article is organized as follows: first, we describe the empirical method for fast protein pK_a prediction and the rationale behind the approach. Second, we present the pK_a predictions made with the empirical method and compare them to those made with other methods. Based on the predictions, we analyze the structural determinants of carboxyl pK_a values in proteins. Some interesting pK_a sites in various proteins are discussed in detail. Finally, we summarize our work and discuss future directions.

THEORY

The pK_a value of an ionizable group in a protein is predicted by applying an environmental perturbation,

The Supplementary Materials referred to in this article can be found at www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: the National Science Foundation; Grant number: MCB 0209941; Grant sponsor: the National Institutes of Health; Grant number: GM 46869.

[†]Current address: Department of Chemistry, Iowa State University, Ames, Iowa 50011.

*Correspondence to: Jan H. Jensen, Department of Chemistry and Center for Biocatalysis and Bioprocessing, The University of Iowa, Iowa City, Iowa 52242. E-mail: Jan-Jensen@uiowa.edu

Received 11 March 2005; Revised 12 May 2005; Accepted 2 June 2005

Published online 17 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20660

TABLE I. pK_{Model} Values, Definitions of Center Points, and Radii Used to Compute Desolvation Effects Using Equations 6 and 7

	pK_{Model}	Center points (CT)	R_{Local}
C-ter	3.20	$(r_O+r_{OXT})/2$	4.5
Asp	3.80	$(r_{OD1} + r_{OD2})/2$	4.5
Glu	4.50	$(r_{OE1} + r_{OE2})/2$	4.5
Surface his	6.50	$(r_{CG} + r_{ND} + r_{CE} + r_{NE} + r_{CD})/5$	4.0
Buried his	6.50	$(r_{CG} + r_{ND} + r_{CE} + r_{NE} +r_{CD})/5$	6.0
N-ter	8.00	r_N	4.5
Cys	9.00	r_{SG}	3.5
Tyr	10.00	r_{OH}	3.5
Lys	10.50	r_{NZ}	4.5
Arg	12.50	r_{CZ}	5.0

ΔpK_a, to the unperturbed intrinsic pK_a value of the group, pK_{Model}:

$$pK_a = pK_{Model} + \Delta pK_a \tag{1}$$

Both the pK_{Model} and ΔpK_a values are determined empirically. We use pK_{Model} values similar to those used in other studies[13,16,17] (Table I). For ammonium groups of N-termini and carboxyl groups of C-termini and Asp residues, the substituent effects of the backbone peptide bonds significantly decrease their pK_a values.[21] These effects are assumed to be constants and included in the pK_{Model} values. For example, we use a $pK_{Model} = 3.8$ for Asp, which is lower than the $pK_{Model} = 4.5$ for Glu by 0.7 units.

Next we present the empirical relationships between the ΔpK_a and protein structures. The functional forms of Equations 3–5, 7, and 11 as well as the numerical values of the associated parameters were ultimately chosen based on trial and error. We endeavored to develop the simplest possible model capable of predicting pK_a values within ca 1 pH unit of experiment for a majority of the cases [cf. Fig. 4].

Hydrogen Bonding

Our previous study[22] of the molecular determinants of the pKa values of Asp and Glu residues in the protein Turkey ovomucoid third domain (OMTKY3) identified hydrogen bonding as the prime pKa determinant. For example, Asp7 forms two hydrogen bonds to the side-chain hydroxyl group and the backbone amide proton of Ser9, thus lowering the pK_a of Asp7 to 2.4 compared to a model value of 3.8. Similarly, the pK_a values of Glu19 and Asp27 are significantly lowered, to 3.2 and 2.2, due to two and three hydrogen bonds, respectively. The simplest empirical relationship between the pK_a shift of a carboxyl group and its hydrogen bonds can be established by assuming a constant pK_a shift (C_{HB}) for each hydrogen bond:

$$\Delta pK_{HB} = N_{HB} \cdot C_{HB} \tag{2}$$

Here N_{HB} is the number of hydrogen bonds for the carboxyl group and ΔpK_{HB} is the total pK_a shift due to the hydrogen bonds. For OMTKY3, using $C_{HB} = -0.6$ and $N_{HB} = 2$, 2, and 3, the pK_a values of Asp7, Glu19, and Asp27 are predicted to be 2.6, 3.3, and 2.0; in excellent agreement to the experimental values of 2.4, 3.2, and 2.2. No hydrogen bonds are found for Glu10 and Glu43 in OMTKY3, thus their pK_a values are predicted to be 4.5 (i.e., pK_{Model} values), within 0.4 units of the experimental values of 4.1 and 4.8. This simple approach also works well for several other proteins (data not shown).

However, Equation 2 is not always applicable because hydrogen-bonding strength and, hence, its effects on pK_a shift are distance and angle dependent.[23] For example, the amide N—H· · ·O hydrogen bond is strong if the H· · ·O distance is ~2.0 Å and the angle ∠NHO is ~180°, while it is significantly weaker if the H· · ·O distance is >3.0 Å or the angle ∠NHO is ~90°. Including weaker hydrogen bonds (larger N_{HB}) may lead to an overestimation of ΔpK_{HB}, while excluding weaker hydrogen bonds (smaller N_{HB}) may lead to an underestimation. Furthermore, different types of hydrogen bonding may have significantly different effects on pK_a shifts and need different C_{HB} values.

To include distance and angle corrections for hydrogen-bonding effects we use a distance function (Fig. 2) to describe the pK_a shifts due to side-chain hydrogen bonds (SDC–HB),

$$\Delta pK_{SDC\text{-}HB} = \begin{cases} C_{HB} & \text{if } D \leq d_1 \\ C_{HB} \cdot \dfrac{D - d_2}{d_1 - d_2} & \text{if } d_1 < D < d_2 \\ 0 & \text{if } d_2 \leq D \end{cases} \tag{3}$$

and a distance/angle function to describe the pK_a shifts due to backbone hydrogen bonds (BKB-HB),

$$\Delta pK_{BKB\text{-}HB} = \begin{cases} -\mathrm{Cos}\theta \cdot C_{HB} & \text{if } D \leq d_1, \theta > 90° \\ -\mathrm{Cos}\theta \cdot C_{HB} \cdot \dfrac{D - d_2}{d_1 - d_2} & \text{if } d_1 < D < d_2, \theta > 90° \\ 0 & \text{if } d_2 \leq D, \theta \leq 90° \end{cases} \tag{4}$$

The functional forms of Equations 3 and 4 is the simplest we could think of that describes the weakening of hydrogen bonds with increasing distance and orientation.[23] By comparison to experimental pK_a values (Fig. 4) we found that the orientation dependence is necessary for accurate pK_a predictions in the case of hydrogen bonds between the ionizable groups and the backbone, but not to the side chains. This may be due to a greater conformational freedom of side chains, compared to the backbone, which allows them to obtain an optimum orientation [$\theta = 180°$ and $-\cos(\theta) = 1$ (θ is defined below)] in the majority of the cases.

The variable D is the distance between the atoms in the hydrogen bond. It is defined as the distance between the carboxyl oxygen atoms and the protons for the hydrogen bonds between carboxyl groups and Asn, Gln, Trp, His, Arg side-chain groups and backbone amides. For other hydrogen bonds D is defined as the distance between the carboxyl oxygen atoms and the other heavy atoms (O, S, and N). The parameter d_1 is the optimum distance for hydrogen bonds at which the ΔpK_{HB} is the maximum

TABLE II. Definition of D and Values of C_{HB}, d_1 and d_2 used to compute pK_a-Shifts Due to Hydrogen Bonding Using Equations 3 and 4

	Group	Type	C_{HB}	d_1 (Å)	d_2 (Å)	D (Å)
Carboxyl	Buried COO^-	STR-HB	+1.60	3.5	3.5	Min. of $O^{1,2}$–$O^{1,2}$
	COO^-	SDC-HB	+0.80	2.5	3.5	Min. of $O^{1,2}$–$O^{1,2}$
	COOH	SDC-HB	−0.80	2.5	3.5	Min. of $O^{1,2}$–$O^{1,2}$
	Ser/Thr–OH	SDC-HB	−0.80	3.0	4.0	Min. of O–$O^{1,2}$
	Gln/Asn/Trp–NH	SDC-HB	−0.80	2.0	3.0	Min. of $H^{1,2}$–$O^{1,2}$
	His–NH	SDC-HB	−0.80	2.0	3.0	Min. of $H^{1,2}$–$O^{1,2}$
	Cys–SH	SDC-HB	−0.80	3.0	4.0	Min. of S–$O^{1,2}$
	Tyr–OH	SDC-HB	−0.80	3.0	4.0	Min. of O–$O^{1,2}$
	Lys–NH	SDC-HB	−0.80	3.0	4.0	Min. of N–$O^{1,2}$
	Arg–NH	SDC-HB	−0.80	2.0	4.0	Min. of $H^{1,5}$–$O^{1,2}$
	Arg–NH	DBL-HB	−1.20	2.2	2.2	Min. of $H^{1,5}$–$O^{1,2}$
	N-ter–NH	SDC-HB	−1.20	3.0	4.5	Min. of N–$O^{1,2}$
	Backbone–NH	BKB-HB	−1.20	2.0	3.5	Min. of H–$O^{1,2}$
	Buried His–NH	STR-HB	−1.60	3.0	3.0	Min. of $H^{1,2}$–$O^{1,2}$
	Buried COOH	STR-HB	−1.60	3.5	3.5	Min. of $O^{1,2}$–$O^{1,2}$
His	Buried Cys–S^-	STR-HB	+3.60	4.0	4.0	Min. of S–$H^{1,2}$
	Buried COO^-	STR-HB	+1.60	3.0	3.0	Min. of $O^{1,2}$–$H^{1,2}$
	Cys–S^-	SDC-HB	+1.60	3.0	4.0	Min. of S–$H^{1,2}$
	Backbone–CO	BKB-HB	+1.20	2.0	3.5	Min. of S–$H^{1,2}$
	Asn/Gln–CO	SDC-HB	+0.80	2.0	3.0	Min. of O–$H^{1,2}$
	COO^-	SDC-HB	+0.80	2.0	3.0	Min. of $O^{1,2}$–$H^{1,2}$
Cys	Buried Cys–S^-	STR-HB	+3.60	5.0	5.0	S–S
	Cys–S^-	SDC-HB	+1.60	3.0	5.0	S–S
	COO^-	SDC-HB	+0.80	3.0	4.0	Min. of $O^{1,2}$–S
	Cys–SH	SDC-HB	−1.60	3.0	5.0	S–S
	Ser/Thr–OH	SDC-HB	−1.60	3.5	4.5	O–S
	Gln/Asn/Trp–NH	SDC-HB	−1.60	2.5	3.5	H–S
	His–NH	SDC-HB	−1.60	3.0	4.0	Min. of $H^{1,2}$–S
	Tyr–OH	SDC-HB	−1.60	3.5	4.5	O–S
	Lys–NH	SDC-HB	−1.60	3.0	4.0	N–S
	Arg–NH	SDC-HB	−1.60	2.5	4.0	Min. of $H^{1,5}$–S
	Backbone–NH	BKB-HB	−2.40	3.5	4.5	H–S
	N-ter–NH	SDC-HB	−2.40	3.0	4.5	N–S
	Buried His–NH	STR-HB	−3.60	4.0	4.0	Min. of $H^{1,2}$–S
	Buried Cys-SH	STR-HB	−3.60	5.0	5.0	S–S
Tyr	Cys–S^-	SDC-HB	+1.60	3.5	4.5	S–O
	COO^-	SDC-HB	+0.80	3.0	4.0	Min. of $O^{1,2}$–O
	Tyr–O^-	SDC-HB	+0.80	3.5	4.5	O–O
	Tyr–OH	SDC-HB	−0.80	3.5	4.5	O–O
	His–NH	SDC-HB	−0.80	2.0	3.0	Min. of $H^{1,2}$–O
	Ser/Thr–OH	SDC-HB	−0.80	3.5	4.5	O–O
	Gln/Asn/Trp–NH	SDC-HB	−0.80	2.5	3.5	H–O
	Lys–NH	SDC-HB	−0.80	3.0	4.0	N–O
	Arg-NH	SDC-HB	−0.80	2.5	4.0	Min. of $H^{1,5}$–O
	N-ter–NH	SDC-HB	−1.20	3.0	4.5	N–O
	Backbone–NH	BKB-HB	−1.20	3.5	4.5	H–O

value. In general, we select $d_1 = 2.0$ Å if the variable D is defined as the hydrogen-bond length, and $d_1 = 3.0$ Å if the variable D is defined as heavy atom distance. The parameter d_2 is the distance where the hydrogen-bonding strength is effectively zero, and is generally selected to be 1.0 Å larger than the corresponding d_1 value. Both d_1 and d_2 are slightly adjusted to best reproduce the experimental pK_a values (Table II). For backbone hydrogen bonds (BKB-HB), θ is defined as the larger of the two open angles ∠NHO.

As discussed above, a $C_{HB} = -0.6$ in Equation 2 can be used for many cases. However, to compensate for the distance/angle corrections and desolvation effects (discussed in next section), a larger value is always used for C_{HB} in Equations 3 and 4. We found $C_{HB} = -0.80$ gives the best agreement with experiment for hydrogen bonds between carboxyl groups and proton-donor side-chain groups. The hydrogen bond between two carboxyl groups will cause the higher "intrinsic" pK_a value to be shifted up and the other down, so $C_{HB} = \pm 0.80$ is used (see Results and Discussion, Asp25 of HIV-1 protease for an example; in general, interactions between ionizable residues are treated in an iterative fashion as described in the Computational Methodology section that follows). $C_{HB} = -1.20$ is used for

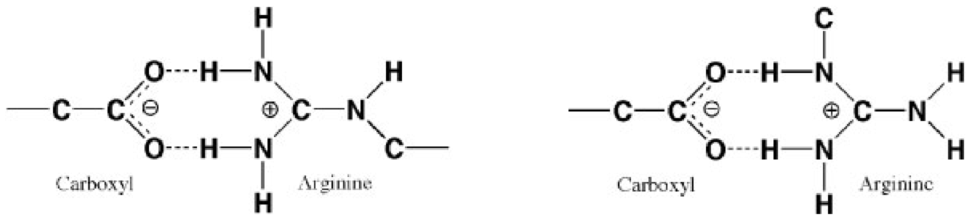

Fig. 1. Schematic represenation of double hydrogen bonds between carboxyl groups and arginine residues.

TABLE III. The Maximum Deviation and RMSD of 83 pK_a Values Predicted with Various Methods for Five Proteins

Protein	Number of pK_a values	PROPKA	Mehler et al.[13]	Nielsen et al.[17]	Demchuk et al.[16]	Georgescu et al.[86]	Wisz et al.[40]
				Maximum deviation			
OMTKY3	13	1.1	1.7	2.6		2.9	1.2
BPTI	14	1.5	0.9	2.0	2.1	1.3	1.0
HEWL	19	1.7	1.3	3.5	1.7	1.5	1.7
RNase A[a]	14	1.2	1.5	2.4	1.7	2.8	1.3
RNase H	23	1.8	1.7	2.5		4.3	1.8
Total	83	1.8	1.7	3.5	2.1	4.3	1.8
				RMSD			
OMTKY3	13	0.44	0.86	1.19		1.26	
BPTI	14	0.60	0.38	0.65	0.66	0.67	0.45
HEWL	19	0.66	0.67	1.11	0.65	0.81	0.69
RNase A[a]	14	0.67	0.66	1.03	0.70	1.04	0.58
RNase H	23	0.72	0.56	1.05		1.37	0.66
Total	83	0.64	0.64	1.03	0.67	1.09	0.57

[a] PROPKA uses the 1RNZ structure.

hydrogen bonds between carboxyl groups and backbone amides, N-termini and Arg residues that form double hydrogen bonds (Fig. 1) to the carboxyl group. Both are considered to be "rigid" hydrogen bonds and are empirically found to result in larger pK_a shifts than other hydrogen bonds.

Strong, low-barrier hydrogen bonds (STR-HBs) may be found between two neighboring groups that are buried and have similar pK_a values (e.g. Asp—Asp, Asp—Glu, Glu—His). We find that for STR-HBs between carboxyl groups and neighboring carboxyl groups and histidine residues, a value of $C_{HB} = \pm 1.6$ gives better agreement with experiment (Table II; this issue is discussed further in the section Results and Discussion, Asp25 of HIV-1 protease). This correction is only applied if the ionizable groups are within a certain distance d_1 (indicated as $d_1 = d_2$ in Table III). The buried groups are identified according to the criteria presented in the next section.

As we will show below the empirical rules in Equations 3 and 4 and the parameters in Table II have been developed and used to successfully predict hundreds of carboxyl pK_a values in more than two dozens of proteins. Next we extend the empirical rules for carboxyl pKa values to those of His, Cys and Tyr residues.

The extension is straightforward. Equations 3 and 4 are used with the C_{HB}, D, d_1 and d_2 values given in Table II. For backbone amide hydrogen bonds to Cys and Tyr residues, θ is the angle $\angle NHX$ (X = S or O). For backbone carbonyl hydrogen bonds to His residues, θ is the larger one of the angles $\angle COH$. Significantly larger C_{HB} values for hydrogen bonds involving Cys thiolates are necessary to reproduce experimental pK_a values for Cys residues and some His residues that are in hydrogen bonding with Cys residues. This is presumably due to the higher charge density on the Cys thiolate group compared to Asp/Glu carboxylates, His imidazolium and Tyr phenol groups.

A His residue has two possible base forms and the hydrogen-bonding effects on its pK_a shifts are more complicated than for carboxyl groups, Cys, and Tyr residues. A Lys or Arg residue has multiple protons and can form hydrogen bonds before and after ionization, so the hydrogen-bonding effects on their pK_a shifts are also complicated. Our tests show that empirical determination of $\Delta pK_{SDC\text{-}HB}$ is relatively difficult for His, Lys, and Arg residues. The experimental Lys pK_a values in proteins are in a relatively narrow range, suggesting that protein interactions such as hydrogen bonding have little effects on them. The pK_a values of Arg residues are usually very high and immeasurable, and limited experimental results are available to derive empirical rules. We found the best agreement with experiment if we only consider the hydrogen bonds to carboxylates, thiolates, Asn/Gln side-chain carbonyl and backbone carbonyl groups when computing pKa values for His residues (Table II). Hydrogen-bonding effects are not considered for Lys and Arg residues. We note that PROPKA predicts the pKa value for the ionizable residue as a whole and not for individual atoms, such as the $N^{\epsilon 2}$ and $N^{\delta 1}$ atoms in His residues or $O^{\delta 1}$ and $O^{\delta 2}$ atom in Asp residues. Accordingly, a specific tautomeric state of neutral His or a specific proton position for protonated Asp/Glu is not assigned as part of the pKa prediction.

Desolvation Effects

Carboxyl groups in the protein interior (i.e., "buried residues") often exhibit pK_a values that are higher than the pK_{Model} values. For example, Glu35 in HEWL (2LZT[24]), Asp10 in RNase H (2RN2[25]), Glu172 in xylananse (1XNB[26]), Asp25 in HIV-1 protease (1HPX[27]) and Asp26 in human thioredoxin (1ERU[28]), which have pK_a values of 6.1, 6.1, 6.7, >6.2 and 8.1, respectively, are all buried residues. Similarly, buried histidine residues that exhibit pK_a values lower than the pK_{Model} values are also found. Typical examples are His149 in xylanase (1XNB[26]), and His61 and His81 in phosphatidylinositol (1GYM), which have pK_a values of <2.3, <3.0, and <3.0, respectively.

Desolvation is the primary reason for these pKa shifts. For C-termini, Asp, Glu, Cys, and Tyr residues, desolvation preferentially increases the energies of the negatively charged base forms thus increasing their pK_a values. For N-termini, His, Lys, and Arg residues, desolvation preferentially increases the energies of the positively charged acid forms and decreases their pK_a values.

Obviously, the desolvation effects on pK_a shifts depend on the degree of protein burial. To establish an empirical relationship that quantitatively links the desolvation effects to protein structures, we need a measure of the degree of protein burial. For instance, the solvent accessible surface (SAS) and the "depth of burial" (i.e., the distance of a group from the protein surface) are two measures commonly used for this purpose. In this study, we use the number of protein atoms around an ionizable group as an easily computable indicator of the degree of burial for the group, and establish the empirical relationship between the desolvation pK_a shifts and the number of these protein atoms.

Two different kinds of desolvation are considered. First, we consider the region within 4 ~ 5 Å to an ionizable group. If there are no protein atoms in this region water molecules presumably occupy it, and the group has the maximum solvent accessible surface (SAS) and solvation energy. If there are some protein atoms in this region less water molecules presumably occupy it, and the group has a lower SAS, and is desolvated. In general, the presence of more protein atoms in this region is taken to mean less water molecules and a larger desolvation effect. Since this desolvation comes from local protein atoms around the ionizable groups, it is referred to as local desolvation.

The simplest empirical relationship between the local desolvation pK_a shifts and the local protein atoms is a constant pK_a shift (C_{Local}) for each nonhydrogen protein atom:

$$\Delta pK_{LocalDes} = N_{Local} \cdot C_{Local} \qquad (5)$$

Here N_{Local} is the number of nonhydrogen atoms within a distance R_{Local} to the center point of the ionizable group (Table I), $\Delta pK_{LocalDes}$ is the pK_a shift due to the local desolvation effects. The atoms of the considered ionizable residue are excluded from N_{Local}. For carboxyl groups, we find an $R_{Local} = 4.5$ and a $C_{Local} = +0.07$ produce the best results. For other ionizable groups, various R_{Local} are used (Table I) but C_{Local} remains ±0.07. All local nonhydrogen protein atoms, i.e., C, N, O and S give rise to the same desolvation effect. Using more sophisticated rules such as atom-specific terms did not improve the final results significantly.

We find that the combined use of Equations 3–5 results in good pK_a predictions for most residues near the protein surface, but works less well for proteins deeper in the protein interior. This observation led us to consider a second desolvation effect due to protein atoms further away from the ionizable groups. Empirically we find that the number of nonhydrogen protein atoms within 15.5 Å of an ionizable group is a good indicator of the degree of bulk burial of the group. Formally, we define an ionizable group as "buried" (otherwise "surface") if

$$N_{15.5Å} \geq 400 \qquad (6)$$

Here $N_{15.5Å}$ is the number of nonhydrogen protein atoms within 15.5 Å of the center of the ionizable group (see Table I for definitions). The atoms of the ionizable residue under consideration are excluded from $N_{15.5Å}$. According to this criterion, Arg29, Lys91, His124, and Glu135 in RNase H (2RN2[25]) are typical "surface" residues that have $N_{15.5Å}$ of 182, 160, 143, and 227, respectively, while Asp10, Arg46, Glu48, and Tyr73 in the same protein are typical "buried" residues that have $N_{15.5Å}$ of 512, 451, 546, and 518, respectively.

As in the case of local desolvation, we establish a simple empirical relationship between the global desolvation pK_a shifts and the number of "excess" protein atoms by assuming a constant pK_a shift (C_{Global}) for each nonhydrogen protein atom for buried residues (cf. Equation 6) only:

$$\Delta pK_{GlobalDes} = (N_{15.5Å} - 400) \cdot C_{Global} \qquad (7)$$

Here $(N_{15.5Å} - 400)$ is the "excess" number of protein atoms, and $\Delta pK_{GlobalDes}$ is the pK_a shift due to the global desolvation. Empirically we find a $C_{Global} = \pm 0.01$ results in the best agreement with experiment.

The total desolvation effects on the pK_a shift thus is:

$$\Delta pK_{Des} = \Delta pK_{GlobalDes} + \Delta pK_{LocalDes} \qquad (8)$$

We use a similar criterion to define a "buried pair" of ionizable groups, which is used to determine strong hydrogen bonding as described in the previous section and the charge–charge interactions described in the next section. Formally, a buried pair is defined as two neighboring groups, 1 and 2, for which:

$$N_{15.5Å}(1) \geq 400 \text{ and } N_{15.5Å}(2) \geq 400 \qquad (9)$$

or

$$[N_{15.5Å}(1) + N_{15.5Å}(2)] \geq 900 \qquad (10)$$

The purpose of Equation 10 is to include the groups that are not buried (e.g., $N_{15.5Å} = 350$) but have a neighboring group that is deeply buried (e.g., $N_{15.5Å} > 550$).

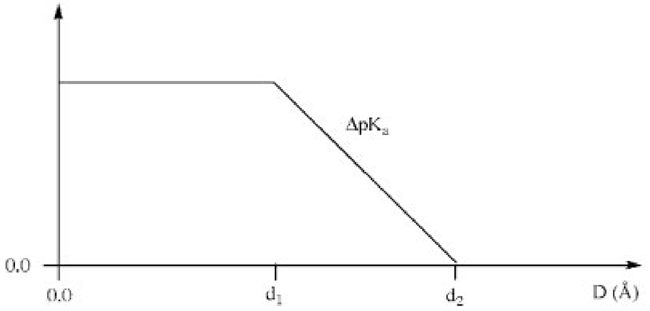

Fig. 2. Schematic representation of the distance dependence of ΔpK$_a$ used in Equations 3, 4, and 11.

Charge–Charge Interactions

The charge–charge interactions between ionizable groups are traditionally believed to be the main determinant of protein pK$_a$ values. For example, one of the first theoretical analyses[29] of protein pK$_a$ values included only interactions between charged residues, and most site-directed mutagenesis studies of pK$_a$ determinants have focused on the effects of charged groups.[19,20,30–39] However, using the empirical rules in Equations 3–5 and 7, we find that the experimental pK$_a$ values of more than 90% of 232 Asp and Glu residues in 26 proteins can be reproduced very well, and the main determinants of their pK$_a$ values are hydrogen bonds to neighboring residues or backbone amides (as discussed in detail below). Only for about ten Asp and Glu residues in 26 proteins, and for some His and Cys residues in these and other proteins, were the pK$_a$ values predicted using Equations 3–5 and 7 significantly different from experiment. The crystal structures show that these residues all have one or more neighboring charged groups, suggesting that the errors are due to the strong Coulomb interactions from these charged groups. Invariably, these residues and their neighboring charged groups are classified as buried. This led us to make the general conclusion that only buried ionizable residues have significant charge–charge interactions. This conclusion is consistent with the use of residue-specific dielectric constants in other pK$_a$ prediction methods.[13,16,17,40] Solvent screening is probably the main reason that surface ionizable residues do not appear to experience significant charge–charge interactions.

Based on the above considerations, we establish a very simple empirical relationship between protein pK$_a$ shifts and the charge–charge interactions between *buried residues* using the same distance function used for hydrogen-bonding effects (Fig. 2 and Equation 3):

$$\Delta pK_{chgchg} = \begin{cases} C_{chgchg} & \text{if } D \leq d_1 \\ C_{chgchg} \cdot \dfrac{D - d_2}{d_1 - d_2} & \text{if } d_1 < D < d_2 \\ 0 & \text{if } d_2 \leq D \end{cases} \qquad (11)$$

Here ΔpK$_{chgchg}$ is the pK$_a$ shift due to the charge–charge interactions, C$_{chgchg}$ is the maximum pK$_a$ shifts due to charge–charge interactions, and D is the distance between the center points of the two ionizable groups (Table I). The parameters d$_1$ and d$_2$ are the distances between which the charge–charge interaction effect is maximized and zero, respectively. The charges of the groups are assumed to be ±1 so they are not explicitly included in the formula. Empirically we find $C_{chgchg} = \pm 2.4$, $d_1 = 4.0$ Å and $d_2 = 7.0$ Å for all residue types produce the best results.

As discussed above, Equation 11 is only applied to buried pairs as defined by Equations 9 or 10. One exception is for carboxyl-Tyr pairs, which are common in proteins (e.g., Asp27-Tyr31 in OMTKY3, 1PPF[41]). Empirically we find that the pK$_a$ values of Tyr residues are always higher than the pK$_{Model}$ value by ~2 units if their hydroxyl groups are in contact with Asp or Glu carboxyl groups, no matter whether they are defined as “buried” or “surface.” These pK$_a$ shifts can be best predicted by including the charge–charge interaction from the carboxylate groups (COO^-) that destabilize the anionic form of Tyr. Excluding this term for “surface” Tyr residues leads to ~2 units errors in their pK$_a$ values. This observation suggests incomplete screening of the charge in the unprotonated form of Tyr residues on the protein surface and may be due to the partially hydrophobic character of the Tyr side chain.

Physical Meanings of the Parameters

In general, there is a clear connection between a pKa change and the change in the deprotonation free energy ($\Delta G = 1.36\Delta pK_a$, in kcal/mol). Thus, for example, $C_{HB} = -0.8$ (the maximum Asp pKa shift due to a Ser residue) can be taken to mean that the Asp–Ser hydrogen bond is on average 1.1 kcal/mol stronger when Asp is ionized than when it is protonated (0.8 is lowered by $0.07 - 0.21$ due to local desolvation, depending on the hydrogen-bond geometry). This number is consistent with our quantum mechanical-based studies of the pK$_a$ values of carboxyl groups in turkey ovomucoid third domain.[42] For comparison, the experimental pK$_a$ values of CH_3CH_2COOH and $HOCH_2CH_2COOH$ are 4.9 and 4.5, respectively. Very similar considerations apply to C$_{chgchg}$.

The desolvation term is a little different. pK$_{Model}$ is often taken to be function of the gas phase deprotonation energy of the ionizable group and the difference in solvation energy of the charged and neutral form of the group. Therefore, 1.36ΔpK$_{Des}$ represents the *change* in the relative solvation energies of the charged and neutral form of the ionizable residue due to the chemical environment. The dominant contribution is likely the change in the solvation energy of the charged form of the ionizable residue.

COMPUTATIONAL METHODOLOGY

Computer Program: PROPKA

A FORTRAN program named PROPKA has been written that calculates pK$_a$ values of all ionizable groups according to the empirical rules described above based on a protein data bank (PDB) file. Because some charge–charge interactions and hydrogen-bonding interactions depend on the charge states of the ionizable groups, which are pK$_a$ and pH dependent, an iterative self-consistent calculation must be performed as described below. For simplicity, only two integer–charge states are considered for each ionizable group at pH $>$ pK_a and pH $<$ pK_a. This simplification gives acceptable results.

PROPKA calculates pK_a values in five steps. The first step is to read the PDB file and identify ionizable groups. In the current implementation, only the nonhydrogen protein atoms will be read and used for pK_a calculations in PROPKA. Hetero-atoms such as water molecules, bound ions, and ligands are not considered. If the PDB file contains more than one chain or NMR models, all the chains and models will be treated as a single protein. The desired structure(s) for pK_a predictions can be selected by editing the PDB files.

The second step uses simple algorithms to determine the positions of N—H protons of backbone amides, Asn, Gln, Trp, His, and Arg residues that may be involved in hydrogen bonding. In this algorithm, the H—N bond length is always 1.0 Å. For a backbone amide groups, the proton position is assigned so that the vector H—N is parallel with the C—O bond vector. For Asn, the H^{δ} position is assigned so that the vector H^{δ}–$N^{\delta 2}$ is parallel with the vector C^{γ} —$O^{\delta 1}$, and the $H^{\delta'}$ position is assigned so that the vector $H^{\delta'}$ —$N^{\delta 2}$ is parallel with the vector X—C^{γ} (X is the midpoint of the $N^{\delta 2}$ and $O^{\delta 1}$), and similarly for Gln residues. For Trp, the $H^{\epsilon 1}$ position is assigned so that the vector $H^{\epsilon 1}$—$N^{\epsilon 1}$ is parallel with the vector $N^{\epsilon 1}$—X (X is the midpoint of the $C^{\delta 1}$ and $C^{\epsilon 2}$ of Trp). For His residue, the $H^{\delta 1}$ position is assigned so that the vector $H^{\delta 1}$-$N^{\delta 1}$ is parallel with the vector $N^{\delta 1}$—X (X is the midpoint of the C^{γ} and $C^{\epsilon 1}$), and the $H^{\epsilon 2}$ position is assigned so that the vector $H^{\epsilon 2}$-$N^{\epsilon 2}$ is parallel with the vector $N^{\epsilon 2}$-X (X is the midpoint of the $C^{\delta 2}$ and $C^{\epsilon 1}$). For Arg, the H^{ϵ} position is assigned so that the vector H^{ϵ}—N^{ϵ} is parallel with the vector N^{ϵ}—X (X is the midpoint of the C^{δ} and C^{ζ}), the four H^{η} positions are assigned so that two H^{η}—N^{η} vectors are parallel with the vector C^{ζ}—N^{ϵ}, one $H^{\eta 1}$—$N^{\eta 1}$ vector is parallel with the vector C^{ζ}—$N^{\eta 2}$ and one $H^{\eta 2}$—$N^{\eta 2}$ vector is parallel with the vector C^{ζ}—$N^{\eta 1}$.

The third step is to calculate temporary pK_a values for the ionizable groups using the protonation states of other ionizable residues that can be easily determined. For example, the temporary pK_a values of carboxyl groups are determined with ΔpK_{Des}, $\Delta pK_{SDC\text{-}HB}$ and $\Delta pK_{BKB\text{-}HB}$, as well as ΔpK_{chgchg} from Lys and Arg residues (but not other Asp, Glu, Cys, or His residues), which are assumed to always be positively charged when carboxyl groups titrate. Only His and Cys residues and other carboxyl groups are considered in the iterative procedure. Note that the temporary pK_a values are not the so-called intrinsic pK_a values determined for a residue when all the other residues are in their neutral states.

The fourth step is to iteratively determine the pK_a values. In the case of carboxyl–carboxyl interactions, the residue with the lowest temporary pK_a value will be taken as negative. This negative residue will further increase the pK_a of the other carboxyl residue. The procedure is then repeated to check whether this pK_a increase will change the pK_a of other residues. Usually one to three iterations are required to reach self-consistency. We emphasize that this iterative procedure only applies to the relatively small number of buried or hydrogen-bonded pairs of ionizable residues.

The fifth step is to print out the predicted pK_a values and specific ΔpK_a terms, i.e., $\Delta pK_{GlobalDes}$, $\Delta pK_{LocalDes}$, $\Delta pK_{SDC\text{-}HB}$, $\Delta pK_{BKB\text{-}HB}$ and ΔpK_{chgchg}.

The total computing time for a protein of around 200 residues is typically 2 sec on a Macintosh PowerBook G4.

Many Cys residues in proteins form disulfide bonds and are not ionizable before reduction. PROPKA identifies disulfide bonds using an S–S distance criterion of 2.5 Å. For bonded Cys groups, no pK_a calculation will be performed and trivial values of 99.99 will be assigned. However, for the Cys_1–Xaa_2–Yaa_3–Cys_4 motifs in the thioredoxin family proteins, PROPKA always calculate the pK_a values of the Cys_1 and Cys_4 even when they are oxidized (bonded). This is useful because sometimes a reduced structure is not available for a protein of interest.

As discussed before, no hydrogen-bonding effects are considered for Lys and Arg residues. However, desolvation effects and Coulomb interactions are generally applicable, and are also used to predict pK_a values of Lys and Arg residues.

An Example: Asp102 in RNase H

We use Asp102 in RNase H as an example to illustrate the pK_a prediction with the PROPKA program. Asp102 in RNase H (2RN2[25]) has 443 and 13 heavy atoms, respectively, within 15.5 Å and 4.5 Å of its carboxyl group and it is therefore defined as "buried," with $\Delta pK_{GlobalDes} = +0.43$ and $\Delta pK_{LocalDes} = +0.91$ [Fig. (3a)].

Asp102 forms double hydrogen bonds (DBL-HB) to Arg46, a BKB–HB to its own NH and another BKB–HB to the NH of Leu103, which results in a $\Delta pK_{DBL\text{–}HB} = -2.40$, a $\Delta pK_{BKB\text{–}HB} = -0.48$ and another $\Delta pK_{BKB\text{–}HB} = -0.46$ [Fig. 3(b)].

Asp102 also experiences a charge–charge interaction from the positively charged Arg46 (both Asp102 and Arg46 are buried and are close to each other), which leads to a $\Delta pK_{chgchg} = -2.40$ [Fig. 3(c)]. The carboxyl group of Asp148 is 6.1 Å from Asp102 and should thus be considered as a possible source of ΔpK_{chgchg} for Asp102. Whether Asp148 titrates before or after Asp102 depends on the relative pK_a values. In PROPKA, this is determined by comparing their temporary pK_a values: the residue with a higher temporary pK_a value will feel the Coulomb interaction from the residue with a lower temporary pK_a value (which titrates first) and has an even higher pK_a value.

Combining all the above ΔpK_a terms, a temporary pK_a value can be obtained for Asp102:

$$\begin{aligned} pK_{Temp}(\text{Asp102}) & \\ &= pK_{Model} + \Delta pK_{Des} + \Delta pK_{HB} + \Delta pK_{chgchg} \quad (12) \\ = pK_{Model} &+ \Delta pK_{GlobalDes} + \Delta pK_{LocalDes} \\ &+ \Delta pK_{SDC\text{-}HB} + \Delta pK_{BKB\text{-}HB} + \Delta pK_{chgchg} \\ = 3.80 &+ 0.43 + 0.91 - 2.40 \\ &- 0.48 - 0.46 - 2.40 = -0.60 \end{aligned}$$

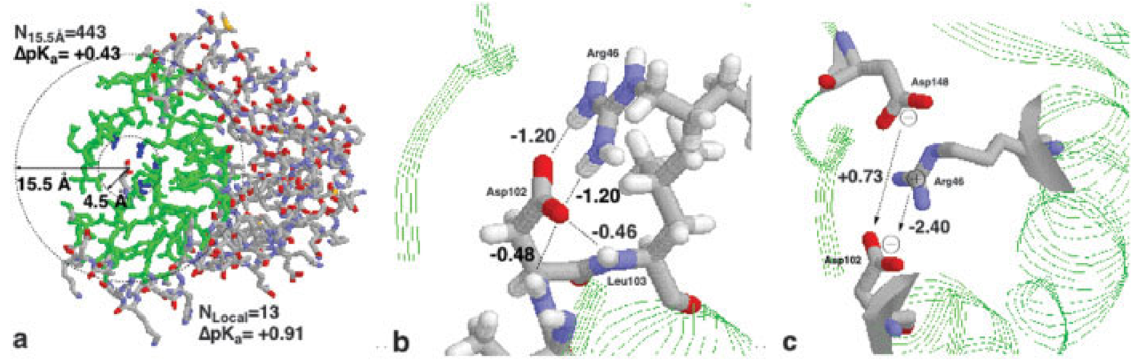

Fig. 3. The determinants of the pK$_a$ value of Asp102 in RNase H (2RN2): (**a**) desolvation effects, (**b**) hydrogen bonding, and (**c**) Coulomb interactions.

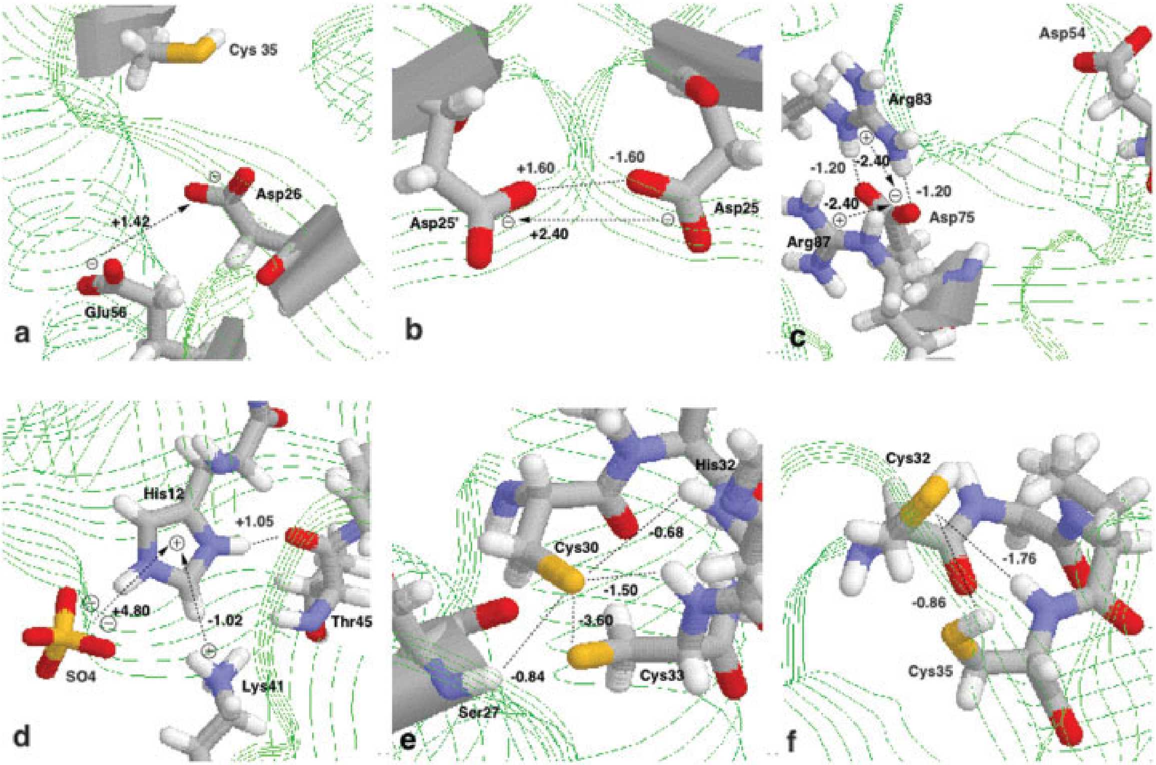

Fig. 5. Some sites with interesting pK$_a$ values, discussed in the Analysis of Select pK$_a$ Value section of the Results and Discussion: (**a**) Asp26 of human thioredoxin, 1ERT; (**b**) Asp25 of chain A and Asp25 of chain B in HIV-1 protease dimmer, 1HPX; (**c**) Asp75 of Barnase, 1A2P chain A; (**d**) His12 of RNase A, 3RN3; (**e**) Cys30 of DSBA, 1DSB chain A; and (**f**) Cys32 of human thioredoxin, 1ERT.

Similar to that of Asp102, the pK$_{\text{Temp}}$ value of Asp148 in RNase H is computed as -0.79. For simplicity, only the two integer-charge states are considered for an ionizable group. Thus when Asp102 titrates, Asp148 is assumed to be fully ionized thus resulting in a $\Delta pK_{\text{chgchg}} = +0.73$ for Asp102 [Fig. 3(c)]. Finally, the pK$_a$ of Asp102 is determined as $+0.13$, while the pK$_a$ of Asp148 is -0.79. The experimental values were reported as <2.0 for both Asp102 and Asp148.[43]

RESULTS AND DISCUSSION

The PROPKA program was used to calculate the pK$_a$ values of ionizable groups in 44 proteins: OMTKY3 (1PPF[41]), ubiquitin (1UBQ[44]), xylanase (1XNB[26]), turkey lysozyme (1LZ3[45]), RNase H (2RN2[25]), B1 protein G (1PGA[46]), B2 protein G (1IGD[47]), barnase (1A2P[48]), BPTI (4PTI[49]), RNase A (3RN3[50] and 1RNZ[51]), HIV-1 protease (1HPX[27], 1HHP[52], 1HVR[53], 1QBR[54], 1QBS[55], 1QBT[54], 1QBU[54], 3HVP[56], and 4HVP[57]), HEWL (2LZT[24]), N-terminus domain of L9 (1DIV[26]), D9K (4ICB[26]), cardiotoxin A5 (1KXI[58]), N-terminal domain of rat CD2 (1CDC[59]), chymotrypsin inhibitor 2 (2CI2[60]), cryptogein (1BEO[61]), oxidized human thioredoxin, (1ERU[28]), reduced human thioredoxin (1ERT[28]), α-Sarcin (1DE3[62] model 1), hirudin (1HIC[63] model 1), BUSI IIA (2BUS[64] model 1), growth factor (1EGF[65] model 1), subunit *c* of H$^+$-transporting

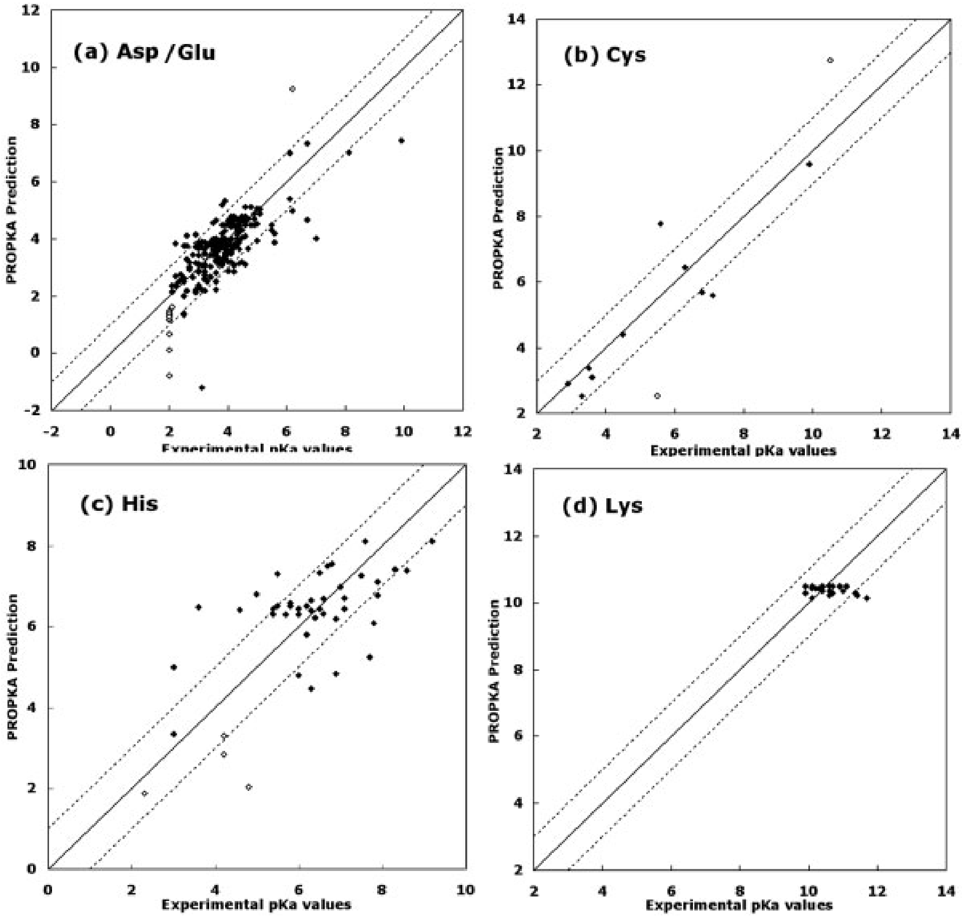

Fig. 4. Correlation plots between experimental and PROPKA predicted pK_a values: (**a**) Asp and Glu, (**b**) Cys, (**c**) His, and (**d**) Lys. The dotted lines represent a pH unit deviation from experiment.

F_1F_0ATP synthase (1A91 model 1), human insulin (1MHI[66] model 1), *Escherichia coli* disulfide oxidoreductase (1DSB[67] chain A), *E. coli* thioredoxin (2TRX[68] chain A), *E. coli* glutaredoxin (1EGO[69] chain A), human muscle creatine kinase (1I0E[70] chain A), papaya proteinase (1MEG[71], 1PPO[72]), human disulfide isomerase (1MEK[73] model 1), serine protease inhibitor (1QLP[74]), rat cathepsin B (1THE[75] chain A), bovine serine proteinase (2TGA[76]), bovine tyrosine phosphatase (1PNT[77]), human cyclophilin A (2CPL[78]), *E. coli* phosphotransferase (1POH[79]), human lysozyme (1LZ1[80]), horse metmyoglobin (1YMB[81]), phosphatidylinositol-specific phospholipase C (1GYM[82]), staphylococcal nuclease (1STG[83]), and cis-trans isomerase (1FKS[84]). Most of these proteins were included in previous survey studies of protein pKa values from which the pKa values used in this study were taken.[2,3,85]

In the following, we report the statistics that reflect the general performance of PROPKA and the details of some PROPKA predictions with comparisons to other methods.

The Accuracy of PROPKA

Comparison to 314 experimental protein pK_a values

The correlation plots between experimental pK_a values and the PROPKA predictions for 233 carboxyl groups, 12 Cys, 45 His, and 24 Lys residues are presented in Figure 4(a–d). Virtually all pK_a values for the Asp, Glu, Cys, His, and Lys residues in the 44 proteins studied in this work are included. The plots show that most of the PROPKA predictions are within 1.0 pH units of the corresponding experimental values. Many points with errors larger than 1.0 are upper or lower limits of experimental values [shown as unfilled marks in Fig. 4(a–d)].

For the 233 carboxyl pK_a values, an RMSD = 0.79 is obtained. Removal of 22 pK_a values that are upper or lower limits and the pK_a of Asp75 in barnase (see discussion below) results in a RMSD = 0.68 for the remaining 210 carboxyl residues. For the 12 Cys pK_a values, a RMSD = 1.39 is obtained. Removal of two values that are of upper or lower limits results in a RMSD = 0.96 for the remaining 10 Cys residues. For the 45 His pK_a values, a RMSD = 1.20 is obtained. Removal of 4 pK_a values that are of upper or lower limits results in a RMSD = 1.15 for the remaining 41 His residues. For the 24 Lys pK_a values, a RMSD = 0.69 is obtained. Reasons for the relatively high RMSD observed for His residues may include misassignment of the ring orientation in the X-ray structure, or the need to consider titrations of each N atom separately. Work on this problem is underway.

Thus for all the 314 pK_a values, PROPKA predictions result in a RMSD = 0.89. Removal of the 29 pK_a values that are of upper or lower limits results in a RMSD = 0.79 for the remaining 285 pK_a values.

Five well-studied proteins

In this section we compare the accuracy of PROPKA to five more conventional pKa prediction methods for five well-studied proteins, OMTKY3, BPTI, HEWL, RNase A, and RNase H, which have been selected as tests or examples in other pK_a-prediction studies.[13,16,17,40,86] Three of the methods are standard Poisson-Boltzmann methods for calculating the pK_a values. In the approach of Demchuck and Wade[16] an energy criterion is used to identify residues as being near the surface or buried in the protein, where after the pK_a calculations are done using respective dielectric constants of 80 and 15. In the approach of Nielsen and Vriend[17] the optimum positions of the protons in hydrogen-bonded networks are determined for each protonation state. In the approach of Georgescu, Alexev, and Gunner[86] different side-chain positions are sampled. The two other approaches[13,40] use distance-dependent screening functions rather than a uniform dielectric to describe solvation as well geometric and electronic polarization. Furthermore, in the approach of Wisz and Hellinga[40] the effect of hydrogen bonding on pK_a values is specifically parameterized.

All methods include parameters (such as a protein dielectric constant or atomic radii) that have been adjusted to obtain agreement with experimental values for some or all of the proteins listed in Table III. The number of parameters that have been adjusted specifically to reproduce experimental pK_a values vary somewhat: the LPBE approaches tend to have fewer parameters that were adjusted manually, while the approach by Wisz and Hellinga has 24 variables which were fitted by a stochastic error minimization. The number of independent variables in PROPKA can be estimated as the number of different values used for a particular parameter, e.g. C_{HB} has five different values: 0.80, 1.20, 1.60, 2.40, and 3.6 (cf. Table II). Counted in this way, PROPKA has 30 parameters, of which 20 are the distance criteria R_{local}, R_{global} = 15.5 Å, d_1, and d_2 used in Equations 3–5, 7, and 11.

The maximum and root mean square deviations (RMSD) from experimental values in the predictions made with PROPKA and other methods are summarized in Table III (the predictions are listed in supplemental materials). The pK_a values for the C-terminus and Tyr31 of OMTKY3, Asp48 and Asp66 of HEWL, Asp14 of RNase A, and Asp102 and Asp148 of RNase H are reported as upper or lower limits. The prediction of the pK_a value of His12 in RNase A is complicated due to the strong Coulomb interaction from anion binding (see below for discussion). These eight residues are not included in the statistics summarized in Table III.

It is clear from the data in Table III that for the five well-studied proteins, the pK_a predictions made by PROPKA are among the best compared to the other methods. A total RMSD of 0.64 (Table III) is obtained for the 83 pK_a values using PROPKA, which is comparable to the accuracy of the methods of Mehler and Guarnieri,[13] Demchuk and Wade,[16] and Wisz and Hellinga;[40] and somewhat better than the methods due to Nielsen and Vriend,[17] and Georgescu, Alexov, and Gunner.[86]

Four proteins not included in the training set

The pK_a values for most of the ionizable residues in four proteins (*Bacillus agaradhaerens* xylanase[87] [BadX, PDB 1QH7], ribonuclease T_1^{88} [Rnase T_1 PDB 1YGW], ribonuclease Sa[20] [Rnase Sa, PDB 1RGG], and bromelain inhibitor VI[89] [BI6, PDB 1BI6]) have been published recently. Since these pK_a values were not used in the training set used to find the optimum parameters in the PROPKA method, they provide important validation of the method. All experimental and predicted values (77 total) are listed in Supplementary Tables S6–S9, and we only summarize the results here. The respective RMSD (maximum) deviations for BadX, Rnase T1, Rnase Sa, and BI6 are 0.69 (1.2), 0.82 (1.9), 0.97 (2.0), and 0.56 (1.7). In computing these values we neglected the pK_a values of Glu17 and Glu178 of BadX, which were in error by 2.6 and 4.5 pH units, respectively. These unusually large errors are likely due to local structural distortions by a ligand (sugar) in the X-ray structure (an X-ray structure of the apo form of BadX is not available). Glu17 is directly hydrogen-bonded to the ligand, while Glu178 is doubly hydrogen-bonded to Arg49, which, in turn, forms two hydrogen bonds to the ligand.

The RMSD values and maximum deviations obtained for these four proteins are comparable to the corresponding values for the five well studies proteins (Table III), suggesting that the PROPKA parameters can be used with equal confidence for proteins outside the training set.

Analysis of 81 interesting pK_a sites

Some of the most meaningful and important uses of protein pK_a predictions are for residues in active sites and structurally important regions in proteins where unusual pK_a values are often found. The ability to predict these pK_a values thus is an important criterion in judging the quality of a method.[90] Table IV lists 81 interesting Asp, Glu, His, and Cys pK_a values (a subset of the 314 pK_a values discussed above) and predictions made with PROPKA and other methods. The "interesting" pK_a values are selected because they (1) are at active sites or (2) at structurally important regions, (3) exhibit large ΔpK_a or (4) errors > 1.5 in PROPKA predictions.

There are 48 interesting Asp and Glu pK_a values in Table IV. For the pK_a values reported as lower or upper limits, the errors are set to be 0.0 if the predictions are within the limits. PROPKA predicts 41 (85%), 36 (75%), and 27 (56%) of them within 1.5, 1.0, and 0.5 pH units of experimental values, respectively. Similarly, the other methods predict these pK_a values to within 1.5 pH units for 82%–87% of the residues, with the exception of the approach of Nielsen and Vriend for which the corresponding percentage is 64%. However, while PROPKA predicts 75% of these interesting pK_a values to within 1 pH unit, only 55%–68% of the pK_a values are predicted with this accuracy by the other methods. Finally, the other methods predict between 24%–50% of the pK_a values to within 0.5 pH units. Since the approach by Demchuk and Wade has only been applied to the prediction of pK_a values for 7 of the 48 residues, we excluded this study in our comparison.

In the case of His and Cys residues we identified 21 and 12 pKa values as interesting (Table IV). PROPKA predicts 11 (52%) and 9 (42%) of the 21 His values, and 10 (83%) and 10 (83%) of the 12 Cys values to within 1.5 and 1.0 of the experimental values, respectively. Unfortunately, the other methods have not been applied to enough of these residues for a meaningful comparison. In general, only very few predictions of Cys pK_a values have appeared in the literature. Dillet et al.[91] reported predictions for the Cys32 and Cys35 of *E. coli* thioredoxin, but both are too high. Foloppe et al.[92] reported predictions for Cys11 and Cys14 of *E. coli* glutaredoxin with good results.

In summary, PROPKA can predict the pK_a values of "interesting" Asp and Glu residues to within 1 pH unit in 75% of the cases considered here compared to 55%–68% for current state-of-the-art pK_a prediction methods involving, for example, extensive fitting to experimental data or multiple side-chain conformations. The better performance of PROPKA, compared to other methods, may reflect the larger number of parameters used in PROPKA that have been adjusted to fit a larger dataset of carboxyl pK_a values.

The corresponding performance of PROPKA for pK_a values of His residues leaves room for improvement (42%) in subsequent versions. Finally, while the accurate prediction of pK_a values of Cys residues presents a significant challenge to current pK_a prediction methods, PROPKA predictions are within 1 pH unit of experimental values in 83% of the cases considered here. A few of the cases in which PROPKA is in error by $>$ 1.5 pH units are discussed next.

Analysis of Select pK_a Values

Asp26 in human thioredoxin

The experimental pKa values of Asp26 in oxidized and reduced thioredoxin are 8.1 and 9.9 pH units.[93] The latter value is among the highest carboxyl pK_a values observed in a protein and is therefore of particular interest. PROPKA predicts corresponding pK_a values of 7.0 and 7.4 pH units, a result of desolvation ($\Delta pK_{GlobalDes}$ = 1.5 and 1.6) and charge–charge interactions with Glu56 (ΔpK_{chgchg} = 1.2 and 1.4). Thus, PROPKA does not reproduce the 1.8-unit difference in pK_a values observed for the oxidized and reduced form of thioredoxin, and the 2.5 unit error in the latter pKa value is one of the largest observed in this study.

The source of the large error in the pK_a value of Asp26 in the reduced form is likely to be the neglect of a charge–charge interaction with Cys35 (ΔpK_{chgchg} = 0.9). PROPKA predicts temporary pK_a values (i.e., values computed without the Cys35–Asp26 interaction) of 7.4 and 7.7 for Asp26 and Cys35, respectively. Since the temporary pK_a value of Asp26 is lower, PROPKA assigns the 0.9 unit shift to the pK_a of Cys35, which increased to 8.6. The 0.3 unit difference in temporary pK_a values is well within the usual errors observed for PROPKA and if the 0.9 unit shift is assigned to the pK_a value of Asp26 instead, the resulting pK_a value of 8.3 is in significantly better agreement with the experimental value of 9.9. Furthermore, since the Cys35–Asp26 charge–charge interaction is absent in the oxidized form of thioredoxin (where a Cys35–Cys32 disulfide bond is formed), the predicted difference in pK_a values becomes 1.3 units,[93] in reasonable agreement with the observed difference of 1.8. Finally, mutating Cys35 to Ala leads to a 1.3-unit decrease in the pK_a value of Asp26,[93] in good agreement with the 0.9-unit interaction predicted by PROPKA.

Asp25 of HIV-1 protease

PROPKA predicts pK_a values of 3.8 and 9.3 (using the 1HPX structure; similar values and determinants are predicted using other structures such as 1HHP,[52] 1HVR,[53] 1QBR,[54] 1QBS,[55] 1QBT,[54] 1QBU,[54] 3HVP,[56] and 4HVP[57]) for the catalytic dyad of HIV protease. The latter is the highest carboxyl pK_a value predicted by PROPKA. Unfortunately, the experimental values are difficult to obtain due to the autolytic breakdown at protein concentrations needed for NMR.[94] An NMR study[95] at pH 5.9 suggests that both pK_a values are $<$ 5.9, but the interpretation is somewhat ambiguous.[94] pK_a values extracted from pH-rate profiles range from 3.3–3.7 and 5.5–6.8, depending on the substrate.[96,97]

The PROPKA prediction of 9.3 thus seems to be an overestimation, but it is not clear by how much. Asp25 and Asp25′ in the dyad have temporary pK_a values of 5.37 and 5.41 (due to desolvation). These pK_a values are then shifted by −1.6 and +1.6, respectively, since the Asp–Asp interaction is classified as a strong hydrogen bond. The higher pK_a value is then further shifted by 2.3 due to charge–charge interactions. This may be the source of the overestimation since that value assumes that the inter-Asp separation is unchanged upon deprotonation, whereas it may increase due to the charge–charge repulsion. In general, both pKa values should be assigned to the dyad; the assignment of the pKa values to individual residues is only done for computational convenience.

Asp75 of barnase

PROPKA predicts a pK_a value of −1.2 units for Asp75 in barnase, while the experimental value obtained by Fersht and coworkers[98] is 3.1. The 4.3-unit difference is the largest discrepancy between experiment and PROPKA for carboxyl pK_a values. The low PROPKA value is due to a double hydrogen bond ($\Delta pK_{DBL\text{-}HB}$ = −2.40) and charge–charge interaction (ΔpK_{chgchg} = −2.40) with Arg83 and another charge–charge interaction (ΔpK_{chgchg}= −2.40) with Arg87. The prediction is based on a 1.5-Å resolution structure[48] and all three residues are in extensive hydrogen-bonded networks and free of crystal contacts. Thus, the pK_a error is probably not due to an error in structure.

Fersht and coworkers[98] noted evidence of coupling in the titration curves of Asp75 and Asp54, as well as Asp86. While the titration curve of Asp75 is not shown in the paper, the corresponding curve for Asp54 is approximately ∩-shaped, and extracting and assigning an accurate pKa

TABLE IV. Comparison of PROPKA Predictions of Unusual pK_a Sites[a] with Five Other pK_a Prediction Methods Due to Mehler and Guarnieri (MG); Nielsen and Vriend (VG); Demchuck and Wade (DW); Georgescu, Alexov; and Gunner (GAG), and Wisz and Hellinga (WH)

Protein	Residue	Exp.	PropKa	MG[13]	NV[17]	DW[16]	GAG[86]	WH[40]
OMTKY3 (1PPF, 2OVO)	Asp27	2.2	2.4	3.3	3.5		3.3	3.4
	Glu19	3.2	3.1	3.6	2.7		**1.6**	4.2
HEWL (2LZT)	Asp48	<2.5	1.4	3.1	2.7	2.7	1.7	3.1
	Asp66	<2.0	1.3	2.6	2.2	2.9	2.6	2.1
	Glu7	2.7	3.7	3.9	3.1	2.9	2.2	3.3
	Glu35	6.1	5.0	4.9	5.7	**4.5**	6.2	5.2
RNase A (3RN3)	Asp14	<2.0	1.4	2.6	1.8	2.3	0.6	2.4
	Asp121	3.1	3.7	3.9	2.1	2.1	3.2	2.2
	Glu2	2.8	2.7	**4.3**	**0.4**	2.6	**1.3**	3.9
RNase H (2RN2)	Asp10	6.1	7.0	5.4	6.2		**10.4**	
	Asp70	2.6	**4.1**	**4.3**	**4.2**		2.3	
	Asp102	<2.0	0.1	3.2			2.0	
	Asp148	<2.0	−0.8	**4.2**			1.1	
	Glu48	4.4	4.6	4.0	**1.9**		5.5	4.6
	Glu57	3.2	2.6	2.8	**1.5**		2.5	**5.0**
	Glu119	4.1	3.5	3.9	**2.5**		3.1	3.4
	Glu129	3.6	3.5	3.0	**1.7**		2.8	3.4
Xylanase (1XNB)	Asp83	<2.0	1.4		−0.1			
	Asp101	<2.0	1.5		0.4			
	Glu78	4.6	5.1		3.3			4.3
	Glu172	6.7	7.3		**5.0**			**4.7**
Barnase (1A2P)	Asp54	<2.2	2.7		0.3		2.3	
	Asp75	3.1	**−1.2**				4.5	2.4
	Asp93	<2.0	0.7				1.4	
	Asp101	<2.0	1.2		2.4		3.2	
	Glu73	<2.1	1.6				0.6	
Protein G (1PGA)	Glu27	4.5	3.2		**2.1**		3.8	3.7
HIV-1 protease (1HPX)	Asp25	>6.2	9.3					
	Asp25	<2.5	3.8					
CardiotoxinA5 (TKX1)	Asp59	<2.3	2.5					
CD2d1 (1CDC)	Glu41	6.7	**4.7**					
Thioredoxin (Oxid. 1ERU)	Asp26	8.1	7.0					
	Glu6	4.9	4.9					
	Glu56	5.0	5.1					
	Glu68	5.1	4.9					
Thioredoxin (Red. 1ERT)	Asp26	9.9	**7.4**					
	Glu6	4.8	5.1					
	Glu56	5.0	4.5					
	Glu68	4.9	4.7					
α-Sarcin (1DE3, NMR)	Asp41	<3.0	2.9					
	Asp77	<3.0	3.4					
	Asp91	<3.0	3.7					
	Asp102	<3.0	3.7					
	Asp105	<3.0	3.1					
	Glu96	5.1	5.0					4.3
ATP synthase (1A9I, NMR)	Asp7	5.6	**3.9**					
	Asp61	7.0	**4.0**					
Human insulin (1MHI)	Glu13	2.2	**3.8**					**4.3**
Error < 1.5			41/48	14/17	14/22	6/7	20/23	16/19
Error < 1.0			36/48	10/17	12/22	6/7	15/23	13/19
Error < 0.5			27/48	4/17	11/22	4/7	10/23	5/19
OMTKY3 (1PPF, 2OVO)	His52	7.5	7.3	7.4	6.1		8.2	
Xylanase (1XNB)	His149	<2.3	1.9		**3.9**			
BPTP (1PNT)	His66	8.3	7.4					
	His72	9.2	8.1					
HEWL(2LZT)	His15	5.5	**7.3**	6.0	4.9	5.8	6.4	6.1
Cyclophilin A (2CLP)	His54	<4.2	2.9					
	His92	<4.2	3.3					
	His126	6.3	**4.5**					

TABLE IV. (Continued)

Protein	Residue	Exp.	PropKa	MG[13]	NV[17]	DW[16]	GAG[86]	WH[40]
Metmyoglobin (1YMB)	His24	<4.8	2.0					
	His36	7.8	**6.1**					
Phosphatidylinositol	His32	7.6	8.1					
(1GYM)	His61	<3.0	**5.0**					
	His81	<3.0	3.3					
	His82	6.9	**4.8**					
RNase A (3RN3)	His12	5.8	**1.8**	6.3	**2.3**	6.2	6.7	6.0
RNase H (2RN2)	His114	5.0	**6.8**	5.2	5.2		4.5	6.1
Cathepsin B (1THE_A)	His110	6.9	6.2					
	His111	7.7	**5.2**					
	His199	8.6	7.4					
Bovine CT (2TGA)	His40	4.6	**6.4**					
1FKS, NMR	His25	<3.6	**6.5**					
Error < 1.5			11/21	4/4	3/5	2/2	4/4	3/3
Error < 1.0			9/21	4/4	2/5	2/2	4/4	2/3
DSBA (1DSB_A)	Cys30	3.5	3.4					
Human thioredoxin (1ERT)	Cys32	6.3	6.4					
E. coli thioredoxin	Cys32	7.1	**5.6**			**>9.2**[b]		
(2TRX_A)	Cys35	9.9	9.6			**>15**[b]		
E. coli Glutaredoxin	Cys11	<5.5	2.5		4.9[c]			
(1EGO_1)	Cys14	>10.5	12.8		13.4[c]			
HPD1 (1MEK_1)	Cys36	4.5	4.4					
1IUE_A	Cys283	5.6	**7.8**					
1PPO	Cys25	3.3	2.5					
1MEG	Cys25	2.9	2.9					
1THE_A	Cys29	3.6	3.1					
1QLP	Cys232	6.8	5.7					
Error < 1.5			10/12		2/2[c]	0/2[b]		
Error < 1.0			10/12		2/2[c]	0/2[b]		

[a] Predictions with errors ≥ 1.5 are in bold and underlined.
[b] Predictions made by Dillet et al.[91]
[c] Predictions made by Foloppe et al.[92]

value from such data is difficult. Interestingly, the predicted pK_a value of Asp54 (2.7 units) is quite close to the value of 3.1 assigned to Asp75, while the predicted value for Asp75 (−1.3 units) is consistent with the experimental estimate assigned to Asp54 (<2.2 units).

His12 of RNase A

PROPKA predicts a very low pK_a of 1.8 units for His12 in RNase (using the 3RN3) structure, a result primarily of desolvation ($\Delta pK_{GlobalDes} = -2.0$ and $\Delta pK_{LocalDes} = -2.7$). This value is significantly lower than the commonly reported experimental value of pK_a values of 5.8–6.3.[8] Interestingly, previous theoretical studies also predicted low pKa values (2.3–4.6) for this residue.[8,17,86] The discrepancy is usually attributed to ion binding. Virtually all crystal structures of RNase A show a sulfate bound near His12, but chloride ions have also been observed,[51] and the pK_a of His12 is known to be strongly salt-dependent.[8] Since some amount of salt is always present in the experimental pK_a measurements, it is possible that the experimental value reflects titrations in the presence of a phosphate or several Cl^- ions.

Continuum electrostatics calculations with and without PO_4^{2-} have shown pK_a differences of up to 3.9 units, and the presence of the ion generally leads to much better agreement with experiment.[8,86] Similarly, within the PROPKA approach, a −2 charge of sulfate could increase the pK_a of His12 by as much as 4.8 units (twice the current maximum charge–charge perturbation of 2.4), which would increase the predicted pK_a to 6.6. The effect of bound ions on pK_a values will be included in the next version of PROPKA, subject to extensive testing and benchmarking.

The Cys_1-Xaa_2-Yaa_3-Cys_4 structural motif

Thioredoxin family enzymes catalyze the breaking and formation of disulfide bonds in proteins. They contain the Cys_1-Xaa_2-Yaa_3-Cys_4 motif with Cys_1 less buried than Cys_4. When reduced, Cys_1 exhibits a much lower pK_a values than Cys_4, with pK_a values of 3.5, 4.5, <5.5, 6.3, and 7.1 in disulfide bond enzyme A (DsbA, Cys30), human protein disulfide isomerase (hPDI, Cys36), *E. coli* glutaredoxin (GRX, Cys11), human (hTRX, Cys32), and *E. coli* thioredoxin (bTRX, Cys32), respectively.[3]

In DsbA [1DSB, Fig. 5(e)], the motif is Cys30-Pro31-His32-Cys33. Using the 1DSB chain A structure, PROPKA predicts a pK_a value of 3.4, in good agreement with the experimental value of 3.5. According to PROPKA, the low pK_a value is primarily a result of a 3.6-unit decrease due to a strong hydrogen bond interaction with Cys33 (since both are classified as buried residues). The pK_a is further lowered by three hydrogen bonds to the backbone amide groups of Cys33 (−1.5), His32 (−0.7), and Ser27 (−0.8).

In hDPI the motif is Cys36-Gly37-His38-Cys39, and PROPKA predicts a pKa value of 4.4 for Cys36 (using the first NMR model of the 1MEK), which is in good agreement with the experimental value of 4.5. The lowering is predicted to be due to a hydrogen bond interaction with Cys33 (−1.6, since Cys33 is not classified as buried). The pK_a is further lowered by three hydrogen bonds to the backbone amide groups of Cys39 (−0.7), His38 (−1.1), and Ala32 (−1.2). The pK_a value of Cys36 in hPDI is therefore higher than the corresponding pK_a value of Cys30 in DsbA, because of weaker Cys–Cys hydrogen bonding due to a more solvent exposed active site in hPDI.

Conversely, the pK_a of Cys11 in the Cys11-Pro12-Tyr13-Cys14 motif in GRX is predicted to be 2.5 (using the first NMR model in the 1EGO file), which is consistent with the experimental estimate of <5.5. Just as for DsbA, Cys14 is classified as buried and the Cys11–Cys14 thus lowers the pK_a by 3.6 units. As in the two other proteins, the pK_a is lowered further by the backbone hydrogen bonds, but in this case to Cys14 (−0.7), Gly10 (−1.0), and Arg9 (−1.9).

Finally, in the case of *E. coli* and human TRX, PROPKA predicts pK_a values of 6.4 and 5.6 (using chain A of the 2TRX structure and 1ERT, respectively) for Cys32 in their Cys32-Gly33-Pro34-Cys35 motifs. These predictions are, respectively, in good and reasonable agreement with the experimental values of 6.3 and 7.1. According to PROPKA the pK_a values are due to hydrogen-bonding with the side chain and backbone of Cys35, which lowers the pK_a by 1.6 and 1.6 for *E. coli* TRX and 0.9 and 1.8 for human TRX [Fig. 5(f)], respectively. Thus, the pK_a values are higher in the thioredoxins compared to hDPI because of fewer backbone hydrogen bonds. The poorer agreement with experiment for *E. coli* TRX is probably because the structure used for the prediction is that of the oxidized enzyme, which would lead to an underestimation of the S–S bond length (and therefore a larger pK_a shift) compared to the reduced geometry (which is not available). For example, in human thioredoxin the S–S distance in reduced TRX is 3.92 Å compared to 2.02 Å in oxidized TRX.

Thus, the prime determinants of the pK_a values of Cys_1 in the Cys_1–Xaa_2–Yaa_3–Cys_4 structural motif is predicted to be the strength of the Cys_1–Cys_4 hydrogen bond (which is modulated by the solvent accessibility of Cys_4) and the number of hydrogen bonds to backbone amides. These conclusions are in general agreement with other studies, in particular the work by Foloppe, Nilsson, and coworkers.[92] It has also been suggested that helix dipole effects[99] help lower the pK_a values of these residues.[3,100] However, we find that the pK_a values can be satisfactorily predicted without including this effect, in general agreement with the findings of Aqvist and Warshel.[101]

The Determinants of Asp and Glu pK_a

All the 232 Asp and Glu residues with known pK_a values (including the 48 interesting ones) discussed in the previous sections are from 26 proteins. In these 26 proteins there are a total of 269 Asp and Glu residues (some with unknown pK_a values): 134 Asp residues and 135 Glu residues. The statistics of the determinants of these 269 pKa values predicted with PROPKA are listed in Table V.

According to Table V, the hydrogen-bonding effects are the main determinants of Asp and Glu pK_a values in these 26 proteins. In total, backbone-amide hydrogen bonds, side-chain hydrogen bonds and charge–charge interactions appear 197, 221, and 25 times, respectively, for the 269 Asp and Glu residues. Hydrogen bonding contributes to the pK_a shifts of Asp residues roughly twice as often as for Glu residues (272 vs. 146 times). This difference is mainly due to fewer hydrogen bonds between Glu side chains and the protein backbone (145 vs. 52 times), consistent with previous observations for other proteins. As a result, the average pK_a shift is larger in magnitude for Asp residues than for Glu residues, as observed previously in a survey study by Forsyth and Robertson.[2] Furthermore, the average pK_a values of Asp (3.2) and Glu residues (4.2) predicted by PROPKA are in good agreements with the values obtained in the survey study of experimental pKa values: 3.4 and 4.1, respectively.

The average pKa shift induced by hydrogen bonding is roughly 0.5 pH units for both Asp and Glu residues with a standard deviation of about 0.3. Within the standard deviations there is no appreciable difference in the average pK_a shifts induced by hydrogen bonds to side chains and the protein backbone.

The most common side chain–hydrogen bond interactions are with Arg, Lys and Ser/Thr residues, which account for 63% and 65% of all side-chain–Asp and side-chain–Glu interactions. The single most likely side-chain-induced perturbations of Asp and Glu pK_a values are due to Ser/Thr and Lys residues, respectively. pK_a values of Asp residues are more likely to be perturbed by Arg than by Lys residues, while both are roughly equally likely to perturb pK_a values of Glu residues. However, there are a total of 96 and 209 Arg and Lys residues in the 26 proteins used in this study. Thus, about 50% of all Arg residues perturb the pK_a values of Asp and Glu residues, compared to only 20% of Lys residues.

The local desolvation contributions to the pKa values of Asp and Glu residues are very similar (0.4 versus 0.3 and 0.6 versus 0.5 for surface and buried residues, respectively). For surface Asp residues 93% (107/115) have a desolvation contribution compared to 75% (94/126) for surface Glu residues, which is consistent with the observation that Glu forms fewer hydrogen bonds. Interestingly, the average global desolvation contribution is somewhat larger for buried Glu residues compared to Asp residues (1.5 vs. 0.9). Thus, buried Glu residues tend to be "more buried" than Asp residues, though they were observed less often.

Only 10% of Asp and Glu residues are characterized as buried using the PROPKA criteria, with Asp residues being ca. 50% more likely to be buried than Glu residues. Thus, charge–charge interactions are predicted to affect the pK_a values of only 10% of Asp and Glu residues, most

TABLE V. The Number of Occurrences (N), the Average ΔpK$_a$ (Average) and the Standard Deviation of the ΔpK$_a$ (σ) of Neighboring Groups for the 134 Asp and 135 Glu Residues in 26 Proteins[†]

	Asp						Glu					
	115 Surface			19 Buried			126 Surface			9 Buried		
	N	Average	σ	N	Average	σ	N	Average	σ	N	Average	σ
BKB–HB	**128**	−0.49	*0.40*	**17**	−0.45	*0.40*	**49**	−0.72	*0.41*	**3**	−0.62	*0.35*
SDC–HB												
Arg	**21**	−0.49	*0.28*	**6**	−0.63	*0.24*	**17**	−0.60	*0.22*	**3**	−0.35	*0.27*
Lys	**16**	−0.64	*0.23*	**1**	−0.23		**24**	−0.56	*0.23*			
Thr	**20**	−0.63	*0.23*	**3**	−0.69	*0.19*	**6**	−0.47	*0.32*			
Ser	**12**	−0.60	*0.24*	**1**	−0.80		**9**	−0.44	*0.29*	**2**	−0.80	*0.00*
Tyr	**9**	−0.71	*0.32*	**2**	−0.80	*0.00*	**5**	−0.78	*0.04*	**4**	−0.67	*0.27*
Asn	**4**	−0.53	*0.32*	**4**	−0.64	*0.18*	**4**	−0.55	*0.32*	**2**	−0.76	*0.06*
Gln	**5**	−0.59	*0.24*				**5**	−0.51	*0.13*	**1**	−0.80	
Asp	**3**	−0.38	*0.07*	**1**	−1.60							
Asp(+)	**3**	+0.38	*0.07*	**1**	+1.60		**3**	+0.49	*0.12*			
Glu	**3**	−0.49	*0.12*				**1**	−0.41				
Glu(+)							**1**	+0.41				
His	**1**	−0.11		**1**	−0.80		**3**	−0.59	*0.21*			
N+	**3**	−0.73	*0.09*				**2**	−0.46	*0.61*			
Trp	**2**	−0.54	*0.37*									
Arg(DBL-HB)	**1**	−2.40		**4**	−2.40	*0.00*	**2**	−2.40	*0.00*			
Total(−)	**100**	−0.60	*0.30*	**23**	−1.00	*0.71*	**78**	−0.60	*0.38*	**12**	−0.64	*0.25*
Total(+)	**3**	+0.38	*0.07*	**1**	+1.60		**4**	+0.47	*0.11*			
Charge–Charge												
Arg				**7**	−1.91	*0.78*				**4**	−1.06	*0.79*
Asp				**5**	+1.36	*0.84*				**3**	+0.15	*0.19*
Glu				**3**	+0.92	*−0.68*				**1**	+1.07	
His				**1**	−0.15							
Lys										**1**	−1.94	
Total(−)				**8**	−1.69	*0.95*				**5**	−1.23	*0.79*
Total(+)				**8**	+1.19	*0.76*				**4**	+0.38	*0.49*
Desolvation												
Local	**107**	+0.41	*0.28*	**19**	+0.59	*0.24*	**94**	+0.33	*0.23*	**9**	+0.47	*0.27*
Global	**0**			**19**	+0.91	*0.70*	**0**			**9**	+1.47	*0.82*

[†]In these 26 proteins, there are 89 surface and seven buried Arg, 211 surface and eight buried Lys, 71 surface and 24 buried Tyr, 37 surface and five buried His.

by interactions with Arg or other Asp/Glu residues. However, while such interactions are comparatively rare they may be important for catalytic function as discussed next.

Eight of the 26 proteins discussed in this section are enzymes, and six of these utilize Asp or Glu residues as a general acid or base in their catalytic mechanism: Glu96 (base) in α-sarcin,[102] Glu73 (base) in barnase,[103] Asp 25/Asp125 (base) in HIV protease,[95] Glu35 (acid) and Asp52 (base) in lysozyme,[104] Asp70 or Glu48 (base) in RNase H1,[105] and Glu172 (acid) and Glu78 (base) in xylanase.[106] All ten residues are characterized as buried and have pK$_a$ values that are perturbed by charge–charge interactions. Clearly, a more exhaustive study of the determinants of pK$_a$ values of catalytic residues is needed (and underway) to explore the generality of this observation. However, the observation is consistent with the suggestion that catalytic residues tend to have unusual acid/base properties.[5,6]

SUMMARY AND FUTURE DIRECTIONS

The empirical PROPKA method is presented for structure-based prediction and rationalization of protein pK$_a$ values of all ionizable residues in a protein in a matter of seconds. The method can predict pK$_a$ values with an overall RMSD of 0.79 from experimental values. The overall accuracy is comparable to current state-of-the-art protein pK$_a$ prediction methods, such as those based on Poisson-Boltzmann electrostatics, but seems to make better predictions for Asp, Glu, and Cys residues with highly shifted pK$_a$ values.

In the PROPKA approach there are three different kinds of pK$_a$ perturbations: desolvation, hydrogen bonding, and charge–charge interactions. The latter term is only evaluated for pairs of residues that are both classified as being buried within the protein interior. Within the PROPKA approach, hydrogen bonding is the most common source of pK$_a$ perturbations for Asp and Glu residues, while charge–charge interactions only contribute in about 10% of the cases. However, catalytic residues appear to all fall in this category.

The PROPKA program is freely available to the scientific community through a web interface at http://propka.chem.uiowa.edu.

Currently, the PROPKA program ignores possible pK_a shifts due to bound ligands, ions, or water molecules in the protein structure. Work on including these effects is ongoing, as is work on improving the accuracy of predicted pK_a values of His residues.

Finally, while we have emphasized the speed, accuracy, and ease-of-use of the PROPKA approach, perhaps the most important conclusion of this study is that the relationship between a protein structure and the acid/base chemistry of its ionizable groups can be quantitatively understood in terms of a few simple rules.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Science Foundation (MCB 0209941) to JHJ, and a grant from the National Institutes of Health (GM 46869) to ADR. HL gratefully acknowledges a predoctoral fellowship from the Center for Biocatalysis and Bioprocessing at the University of Iowa.

REFERENCES

1. Warshel A. Electrostatic basis of structure-function correlation in proteins. Acc Chem Res 1981;14:284–290.
2. Forsyth WR, Antosiewiez JM, Robertson AD. Empirical relationships between protein structure and carboxyl pK(a) values in proteins. Proteins 2002;48:388–403.
3. Harris TK, Turner GJ. Structural basis of perturbed pK(a) values of catalytic groups in enzyme active sites. IUBMB Life 2002;53: 85–98.
4. Kortemme T, Darby NJ, Creighton TE. Electrostatic interactions in the active site of the N-terminal thioredoxin-like domain of protein disulfide isomerase. Biochemistry 1996;35:14503–14511.
5. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312:885–896.
6. Ondrechen MJ, Clifton JG, Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci USA 2001;98:12473–12478.
7. Nielsen JE, McCammon JA. Calculating pKa values in enzyme active sites. Protein Sci 2003;12:1894–1901.
8. Antosiewicz J, McCammon JA, Gilson MK. The determinants of pK(a)s in proteins. Biochemistry 1996;35:7819–7833.
9. Antosiewicz J, McCammon JA, Gilson MK. Prediction of pH-dependent properties of proteins. J Mol Biol 1994;238:415–436.
10. Bashford D, Karplus M. Pkas of ionizable groups in proteins—atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
11. Ullmann GM, Knapp EW. Electrostatic models for computing protonation and redox equilibria in proteins. Eur Biophys J Biophys Lett 1999;28:533–551.
12. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of Pk(a)S in proteins. Proteins 1993;15:252–265.
13. Mehler EL, Guarnieri F. A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. Biophys J 1999;77:3–22.
14. Antosiewicz J, Briggs JM, Elcock AH, Gilson MK, McCammon JA. Computing ionization states of proteins with a detailed charge model. J Comput Chem 1996;17:1633–1644.
15. Karshikoff A. A simple algorithm for the calculation of multiple-site titration curves. Protein Eng 1995;8:243–248.
16. Demchuk E, Wade RC. Improving the continuum dielectric approach to calculating pK(a)s of ionizable groups in proteins. J Phys Chem 1996;100:17373–17387.
17. Nielsen JE, Vriend G. Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK(a) calculations. Proteins 2001;43:403–412.
18. Forsyth WR, Gilson MK, Antosiewicz J, Jaren OR, Robertson AD. Theoretical and experimental analysis of ionization equilibria in ovomucoid third domain. Biochemistry 1998;37:8643–8652.
19. Forsyth WR, Robertson AD. Insensitivity of perturbed carboxyl pK(a) values in the ovomucoid third domain to charge replacement at a neighboring residue. Biochemistry 2000;39:8067–8072.
20. Laurents DV, Huyghues-Despointes BMP, Bruix M, Thurlkill RL, Schell D, Newsom S, Grimsley GR, Shaw KL, Trevino S, Rico M, et al. Charge-charge interactions are key determinants of the pK values of ionizable groups in ribonuclease Sa (pI = 3.5) and a basic variant (pI = 10.2). J Mol Biol 2003;325:1077–1092.
21. Perrin DD, Dempsey B, Serjeant EP. pKa prediction for organic acids and bases. London; New York: Chapman and Hall; 1981.
22. Li H, Robertson AD, Jensen JH. The determinants of carboxyl pKa values in turkey ovomucoid third domain. Proteins 2004;55: 689–704.
23. Jeffrey GA. An introduction to hydrogen bonding, New York: Oxford University Press; 1997.
24. Ramanadham M, Sieker LC, Jensen LH. Refinement of triclinic lysozyme .2. The method of stereochemically restrained least-squares. Acta Crystallographica B 1990;46:63–69.
25. Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Nakamura H, Ikehara M, Matsuzaki T, Morikawa K. Structural details of ribonuclease-H from Escherichia coli as refined to an atomic resolution. J Mol Biol 1992;223:1029–1052.
26. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
27. Baldwin ET, Bhat TN, Gulnik S, Liu BS, Topol IA, Kiso Y, Mimoto T, Mitsuya H, Erickson JW. Structure of Hiv-1 protease with Kni-272, a tight-binding transition-state analog containing allophenylnorstatine. Structure 1995;3:581–590.
28. Weichsel A, Gasdaska JR, Powis G, Montfort WR. Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. Structure 1996;4:735–751.
29. Tanford C, Kirkwood JG. Theory of protein tritration curves. I. General equations for impenetrable spheres. J Am Chem Soc 1957;79:5333–5339.
30. Sundd M, Iverson N, Ibarra-Molero B, Sanchez-Ruiz JM, Robertson AD. Electrostatic interactions in ubiquitin: stabilization of carboxylates by lysine amino groups. Biochemistry 2002;41:7586–7596.
31. Sundd M, Iverson N, Robertson AD. Investigation of electrostatics in ubiquitin by mutagenesis and NMR. Biophys J 82;2002: 299a.
32. Sundd M, Robertson AD. Rearrangement of charge-charge interactions in variant ubiquitins as detected by double-mutant cycles and NMR. J Mol Biol 2003;332:927–936.
33. Dwyer JJ, Gittis AG, Karp DA, Lattman EE, Spencer DS, Stites WE, Garcia-Moreno B. High apparent dielectric constants in the interior of a protein reflect water penetration. Biophys J 2000;79: 1610–1620.
34. Garcia-Moreno B, Dwyer JJ, Gittis AG, Lattman EE, Spencer DS, Stites WE. Solvent penetration may be responsible for the high dielectric constant inside a protein. Biophys J 1998;74: A132.
35. GarciaMoreno B, Dwyer JJ, Gittis AG, Lattman EE, Spencer DS, Stites WE. Experimental measurement of the effective dielectric in the hydrophobic core of a protein. Biophys Chem 1997;64:211–224.
36. Lee KK, Fitch CA, Garcia-Moreno B. Distance dependence and salt sensitivity of pairwise, coulombic interactions in a protein. Protein Sci 2002;11:1004–1016.
37. Lee KK, Fitch CA, Lecomte JTJ, Garcia-Moreno B. Electrostatic effects in highly charged proteins: salt sensitivity of pK(a) values of histidines in staphylococcal nuclease. Biochemistry 2002;41: 5656–5667.
38. Spencer DS, Weiss D, Stites WE, Garcia-Moreno B, Dwyer JJ, Gittis AG, Lattman EE. The pK(a) of buried ionizable groups in staph nuclease: an experimental measure of the dielectric constant of a protein interior. Biophys J 1998;74:A170–A170.
39. Pace CN, Huyghues-Despointes BMP, Briggs JM, Grimsley GR, Scholtz JM. Charge-charge interactions are the primary determinants of the pK values of the ionizable groups in Ribonuclease T1. Biophys Chem 2002;101:211–219.
40. Wisz MS, Hellinga HW. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51:360–377.

41. Bode W, Wei AZ, Huber R, Meyer E, Travis J, Neumann S. X-ray crystal-structure of the complex of human-leukocyte elastase (pmn elastase) and the 3rd domain of the turkey ovomucoid inhibitor. EMBO J 1986;5:2453–2458.
42. Li H, Robertson AD, Jensen JH. The determinants of carboxyl pKa values in turkey ovomucoid third domain. Proteins 2004;55: 689–704.
43. Oda Y, Yamazaki T, Nagayama K, Kanaya S, Kuroda Y, Nakamura H. Individual ionization-constants of all the carboxyl groups in ribonuclease HI from Escherichia coli determined by NMR. Biochemistry 1994;33:5275–5284.
44. Vijaykumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 1987;194:531–544.
45. Harata K. X-ray structure of monoclinic turkey egg lysozyme at 1.3-Angstrom resolution. Acta Crystallographica D Biol Crystallogr 1993;49:497–504.
46. Gallagher T, Alexander P, Bryan P, Gilliland GL. 2 Crystal-structures of the B1 immunoglobulin-binding domain of Streptococcal protein-G and comparison with NMR. Biochemistry 1994; 33:4721–4729.
47. Derrick JP, Wigley DB. The 3rd Igg-binding domain from Streptococcal-protein-G—an analysis by X-ray crystallography of the structure alone and in a complex with Fab. J Mol Biol 1994;243: 906–918.
48. Mauguen Y, Hartley RW, Dodson EJ, Dodson GG, Bricogne G, Chothia C, Jack A. Molecular-structure of a new family of ribonucleases. Nature 1982;297:162–164.
49. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr B 1983;39:480–490.
50. Howlin B, Moss DS, Harris GW. Segmented anisotropic refinement of bovine ribonuclease-A by the application of the rigid-body TLS model. Acta Crystallogr A 1989;45:851–861.
51. Fedorov AA, JosephMcCarthy D, Fedorov E, Sirakova D, Graf I, Almo SC. Ionic interactions in crystalline bovine pancreatic ribonuclease A. Biochemistry 1996;35:15962–15979.
52. Spinelli S, Liu QZ, Alzari PM, Hirel PH, Poljak RJ. The 3-dimensional structure of the aspartyl protease from the Hiv-1 isolate Bru. Biochimie 1991;73:1391–1396.
53. Lam PYS, Jadhav PK, Eyermann CJ, Hodge CN, Ru Y, Bacheler LT, Meek JL, Otto MJ, Rayner MM, Wong YN, et al. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. Science 1994;263:380–384.
54. Jadhav PK, Ala P, Woerner FJ, Chang CH, Garber SS, Anton ED, Bacheler LT. Cyclic urea amides: HIV-1 protease inhibitors with low nanomolar potency against both wild type and protease inhibitor resistant mutants of HIV. J Med Chem 1997;40:181–191.
55. Lam PYS, Ru Y, Jadhav PK, Aldrich PE, DeLucca GV, Eyermann CJ, Chang CH, Emmett G, Holler ER, Daneker WF, et al. Cyclic HIV protease inhibitors: synthesis, conformational analysis, P2/P2′ structure-activity relationship, and molecular recognition of cyclic ureas. J Med Chem 1996;39:3514–3525.
56. Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SBH. Conserved folding in retroviral proteases—crystal-structure of a synthetic hiv-1 protease. Science 1989;245:616–621.
57. Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, Clawson L, Selk L, Kent SBH, Wlodawer A. Structure of complex of synthetic Hiv-1 protease with a substrate-based inhibitor at 2.3-Aring; resolution. Science 1989;246:1149–1152.
58. Sun YJ, Wu WG, Chiang CM, Hsin AY, Hsiao CD. Crystal structure of cardiotoxin V from Taiwan cobra venom: pH-dependent conformational change and a novel membrane-binding motif identified in the three-finger loops of P-type cardiotoxin. Biochemistry 1997;36:2403–2413.
59. Murray AJ, Lewis SJ, Barclay AN, Brady RL. One sequence, 2 folds—a metastable structure of Cd2. Proc Natl Acad Sci USA 1995;92:7337–7341.
60. McPhalen CA, James MNG. Crystal and molecular-structure of the serine proteinase-ihibitor Ci-2 from barley seeds. Biochemistry 1987;26:261–269.
61. Boissy G, deLaFortelle E, Kahn R, Huet JC, Bricogne G, Pernollet JC, Brunie S. Crystal structure of a fungal elicitor secreted by Phytophthora cryptogea, a member of a novel class of plant necrotic proteins. Structure 1996;4:1429–1439.
62. Perez-Canadillas JM, Santoro J, Campos-Olivas R, Lacadena J, del Pozo AM, Gavilanes JG, Rico M, Bruix M. The highly refined solution structure of the cytotoxic ribonuclease alpha-sarcin reveals the structural requirements for substrate recognition and ribonucleolytic activity. J Mol Biol 2000;299:1061–1073.
63. Szyperski T, Guntert P, Stone SR, Wuthrich K. Nuclear-magnetic-resonance solution. Structure of hirudin(1-51) and comparison with corresponding 3-dimensional structures determined using the complete 65-residue hirudin polypeptide-chain. J Mol Biol 1992;228:1193–1205.
64. Williamson MP, Havel TF, Wuthrich K. Solution conformation of proteinase inhibitor-Iia from bull seminal plasma by H-1 nuclear magnetic-resonance and distance geometry. J Mol Biol 1985;182: 295–315.
65. Montelione GT, Wuthrich K, Burgess AW, Nice EC, Wagner G, Gibson KD, Scheraga HA. Solution structure of murine epidermal growth-factor determined by NMR-spectroscopy and refined by energy minimization with restraints. Biochemistry 1992;31: 236–249.
66. Jorgensen AMM, Kristensen SM, Led JJ, Balschmidt P. 3-Dimensional solution structure of an insulin dimer—a study of the B9(Asp) mutant of human insulin using nuclear-magnetic-resonance, distance geometry and restrained molecular-dynamics. J Mol Biol 1992;227:1146–1163.
67. Martin JL, Bardwell JCA, Kuriyan J. Crystal-structure of the Dsba protein required for disulfide bond formation in-vivo. Nature 1993;365:464–468.
68. Katti SK, Lemaster DM, Eklund H. Crystal-structure of thioredoxin from Escherichia coli at 1.68 Å resolution. J Mol Biol 1990;212:167–184.
69. Xia TH, Bushweller JH, Sodano P, Billeter M, Bjornberg O, Holmgren A, Wuthrich K. NMR structure of oxidized Escherichia coli glutaredoxin—comparison with reduced Escherichia coli glutaredoxin and functionally related proteins. Protein Sci 1992; 1:310–321.
70. Shen YQ, Tang L, Zhou HM, Lin ZJ. Structure of human muscle creatine kinase. Acta Crystallogr D Biol Crystallogr 2001;57: 1196–1200.
71. Katerelos NA, Taylor MAJ, Scott M, Goodenough PW, Pickersgill RW. Crystal structure of a caricain D158E mutant in complex with E-64. FEBS Lett 1996;392:35–39.
72. Pickersgill RW, Rizkallah P, Harris GW, Goodenough PW. Determination of the structure of papaya protease omega. Acta Crystallogr B 1991;47:766–771.
73. Kemmink J, Darby NJ, Dijkstra K, Nilges M, Creighton TE. Structure determination of the N-terminal thioredoxin-like domain of protein disulfide isomerase using multidimensional heteronuclear C-13/N-15 NMR spectroscopy. Biochemistry 1996; 35:7684–7691.
74. Elliott PR, Pei XY, Dafforn TR, Lomas DA. Topography of a 2.0 angstrom structure of alpha(1)-antitrypsin reveals targets for rational drug design to prevent conformational disease. Protein Sci 2000;9:1274–1281.
75. Jia ZC, Hasnain S, Hirama T, Lee X, Mort JS, To R, Huber CP. Crystal-structures of recombinant rat cathepsin-B and a cathepsin B-inhibitor complex—implications for structure-based inhibitor design. J Biol Chem 270;1995:5527–5533.
76. Walter J, Steigemann W, Singh TP, Bartunik H, Bode W, Huber R. On the disordered activation domain in trypsinogen–chemical labeling and low-temperature crystallography. Acta Crystallographica B 1982;38:1462–1472.
77. Zhang M, Vanetten RL, Stauffacher CV. Crystal-structure of bovine heart phosphotyrosyl phosphatase at 2.2-Angstrom resolution. Biochemistry 1994;33:11097–11105.
78. Ke HM. Similarities and differences between human cyclophilin-a and other beta-barrel structures—structural refinement at 1.63 Angstrom resolution. J Mol Biol 1992;228:539–550.
79. Jia ZC, Quail JW, Waygood EB, Delbaere LTJ. The 2.0-Angstrom resolution structure of Escherichia coli histidine-containing phosphocarrier protein Hpr—a redetermination. J Biol Chem 1993; 268:22490–22501.
80. Artymiuk PJ, Blake CCF. Refinement of human lysozyme at 1.5 Å resolution analysis of nonbonded and hydrogen-bond interactions. J Mol Biol 1981;152:737–762.
81. Evans SV, Brayer GD. High-resolution study of the 3-Dimensional structure of horse heart metmyoglobin. J Mol Biol 1990;213: 885–897.

82. Heinz DW, Ryan M, Smith MP, Weaver LH, Keana JFW, Griffith OH. Crystal structure of phosphatidylinositol-specific phospholipase C from Bacillus cereus in complex with glucosaminyl(alpha 1→6)-D-myo-inositol, an essential fragment of GPI anchors. Biochemistry 1996;35:9496–9504.

83. Loll PJ, Quirk S, Lattman EE, Garavito RM. X-Ray crystal-structures of Staphylococcal nuclease complexed with the competitive inhibitor Cobalt(Ii) and Nucleotide. Biochemistry 1995;34: 4316–4324.

84. Michnick SW, Rosen MK, Wandless TJ, Karplus M, Schreiber SL. Solution structure of Fkbp, a rotamase enzyme and receptor for Fk506 and rapamycin. Science 1991;252:836–839.

85. Edgcomb SP, Murphy KP. Variability in the pKa of histidine side-chains correlates with burial within proteins. Proteins 2002; 49:1–6.

86. Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. Biophys J 2002;83:1731–1748.

87. Betz M, Lohr F, Wienk H, Ruterjans H. Long-range nature of the interactions between titratable groups in Bacillus agaradhaerens family 11 xylanase: pH titration of B. agaradhaerens xylanase. Biochemistry 2004;43:5820–5831.

88. Spitzner N, Lohr F, Pfeiffer S, Koumanov A, Karshikoff A, Ruterjans H. Ionization properties of titratable groups in ribonuclease T-1 - I. pK(a) values in the native state determined by two-dimensional heteronuclear NMR spectroscopy. Eur Biophys J Biophys Lett 2001;30:186–197.

89. Hatano K, Kojima M, Tanokura M, Takahashi K. Nuclear magnetic resonance studies on the pK(a) values and interaction of ionizable groups in bromelain inhibitor VI from pineapple stem. Biol Chem 2003;384:93–104.

90. Schutz CN, Warshel A. What axe the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.

91. Dillet V, Dyson HJ, Bashford D. Calculations of electrostatic interactions and pK(a)s in the active site of Escherichia coli thioredoxin. Biochemistry 1998;37:10298–10306.

92. Foloppe N, Sagemark J, Nordstrand K, Berndt KD, Nilsson L. Structure, dynamics and electrostatics of the active site of glutaredoxin 3 from Escherichia coli: comparison with functionally related proteins. J Mol Biol 2001;310:449–470.

93. Qin J, Clore GM, Gronenborn AM. Ionization equilibria for side-chain carboxyl groups in oxidized and reduced human thioredoxin and in the complex with its target peptide from the transcription factor NF kappa B. Biochemistry 1996;35:7–13.

94. Trylska J, Antosiewicz J, Geller M, Hodge CN, Klabe RM, Head MS, Gilson MK. Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. Protein Sci 1999;8:180–195.

95. Smith R, Brereton IM, Chai RY, Kent SBH. Ionization states of the catalytic residues in HIV-1 protease. Nat Struct Biol 1996;3: 946–950.

96. Hyland LJ, Tomaszek TA, Meek TD. Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. Biochemistry 1991;30:8454–8463.

97. Ido E, Han HP, Kezdy FJ, Tang J. Kinetic-studies of human-immunodeficiency-virus type-1 protease and its active-site hydrogen-bond mutant A28s. J Biol Chem 1991;266:24359–24366.

98. Oliveberg M, Arcus VL, Fersht AR. Pk(a) values of carboxyl groups in the native and denatured states of barnase—the Pk(a) values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry 1995;34:9424–9433.

99. Wada A. The alpha-helix as an eletric macrodipole. Adv Biophys 1976;9:1–63.

100. Kortemme T, Creighton TE. Ionization of cysteine residues at the termini of model alpha-helical peptides—relevance to unusual thiol Pk(a) values in proteins of the thioredoxin family. J Mol Biol 1995;253:799–812.

101. Aqvist J, Luecke H, Quiocho FA, Warshel A. Dipoles localized at helix termini of proteins stabilize charges. Proc Natl Acad Sci USA 1991;88:2026–2030.

102. Perez-Canadillas JM, Campos-Olivas R, Lacadena J, del Pozo AM, Gavilanes JG, Santoro J, Rico M, Bruix M. Characterization of pK(a) values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between electrostatic interactions, structure, and catalytic function. Biochemistry 1998;37: 15865–15876.

103. Schreiber G, Frisch C, Fersht AR. The role of Glu73 of barnase in catalysis and the binding of barstar. J Mol Biol 1997;270:111–122.

104. Nielsen JE, McCammon JA. On the evaluation and optimization of protein X-ray structures for pKa calculations. Protein Sci 2003;12:313–326.

105. Babu CS, Dudev T, Casareno R, Cowan JA, Lim C. A combined experimental and theoretical study of divalent metal ion selectivity and function in proteins: application to E. coli ribonuclease H1. J Am Chem Soc 2003;125:9318–9328.

106. Joshi MD, Sidhu G, Nielsen JE, Brayer GD, Withers SG, McIntosh LP. Dissecting the electrostatic interactions and pH-dependent activity of a family 11 glycosidase. Biochemistry 2001;40:10115–10139.