WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

# A statistical approach to the prediction of p$K_a$ values in proteins

**Yun He,[1,3] Jialin Xu,[2] and Xian-Ming Pan[1,2]***

[1] National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

[2] Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China

[3] Graduate University of Chinese Academy of Sciences, Beijing 100049, China

*ABSTRACT*

***We propose a simple model for the calculation of p$K_a$ values of ionizable residues in proteins. It is based on the premise that the p$K_a$ shift of ionizable residues is linearly correlated to the interaction between a particular residue and the local environment created by the surrounding residues. Despite its simplicity, the model displays good prediction performance. Under the sixfold cross test prediction over a data set of 405 experimental p$K_a$ values in 73 protein chains with known structures, the root-mean-square deviation (RMSD) between the experimental and calculated p$K_a$ was found to be 0.77. The accuracy of this model increases with increasing size of the data set: the RMSD is 0.609 for glutamate (the largest data set with 141 sites) and ~1 pH unit for lysine, with a data set containing 45 sites.***

Proteins 2007; 69:75–82.


***Key words: protein ionizable residues p$K_a$ prediction; residue interaction; protonation; deprotonation; electrostatic interaction.***

## INTRODUCTION

Ionizable residues play key roles in biological processes involving proteins, including ligand binding, protein–protein interaction, and protein folding and unfolding. For example, the structure and function of many proteins are dependent on the protonation equilibrium of ionizable residues. Most enzymes perform catalysis with the assistance of ionizable residues that either act directly as acids, bases, and ligands, or less directly through effects on the electrostatic potential at or near the active site.[1] The distribution of surface-charged residues is critical for protein–protein association.[2] Many proteins denature when the pH is changed to very low or very high values. The primary reason for denaturation at extreme pH is that proteins usually have buried residues with highly perturbed p$K_a$ values.[3]

Knowledge of the p$K_a$ values of ionizable residues enables the prediction of the protonation state at a given pH, which is often essential in the prediction of protein properties.[4] However, it is also well recognized that structural and environmental features of the protein can influence the protonation state. For example, p$K_a$ values of solvent-exposed carboxylic residues show little variation and have relatively narrow distributions, while p$K_a$ values for buried residues can range from 2 to 6.7.[5] Considering these factors, it is not surprising that many experimentally determined p$K_a$ values are very different from those derived from model compounds in solution. These effects are referred to as perturbations of the p$K_a$ values, and present a significant barrier to accurate prediction of protein properties.

The p$K_a$ values of ionizable residues can be determined experimentally by nuclear magnetic resonance (NMR), but the applicability of this technique is limited by protein size and solubility.[6–10] On the other hand, there has been a great deal of activity in the fields of computational biology and computational chemistry aimed at developing theoretical methods for calculation of p$K_a$ values in proteins. Most of these theoretical methods employ various treatments to give quantitative descriptions of electrostatic effects, such as free-energy perturbation (FEP) approaches,[11] protein dipole Langevin dipole (PDLD) and PDLD/S-LRA approaches,[12,13] Tanford Kirkwood (TK)[14] and modified TK models,[15] as well as popularly applied Poisson Boltzmann (PB) approaches,[16] which have been described a lot in reviews[4,12,16–19] and other articles.[11,13–15,20–33] These methods have given encouraging results that could provide useful insights into the structure–function correlator for proteins. However, the

The web server that implements our statistical model for the prediction of p$K_a$ values in proteins is freely accessible at http://spg.biosci.tsinghua.edu.cn/~jarod/pkacal/submitjob.cgi.

Grant sponsor: Natural Science Foundation of China; Grant numbers: 30230100, 30670420.

*Correspondence to: Xian-Ming Pan, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China. E-mail: pan-xm@mail.tsinghua.edu.cn

Received 25 December 2006; Revised 7 February 2007; Accepted 20 February 2007

Published online 22 June 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21478

results of macroscopic models depend critically on the choice of protein dielectric constant whose concept is problematic, while the microscopic models are entirely practical but require professional treatment, which has been provided by some research groups.[4,19]

Nowadays, with an increasing number of experimentally available p$K_a$ values for proteins with known structure, prediction of p$K_a$ values by means of the empirical parameterization of the protein p$K_a$ database becomes feasible.[5,34–36] These empirical methods yield a root-mean-square deviation (RMSD) of ~1 pH unit. This performance is comparable to theoretical models, but is achieved with greater speed and ease of use. Wisz and Hellinga[34] modeled solvent and interior effects using multiple dielectric constants obtained by fitting to experimental p$K_a$ values. Godoy-Ruiz *et al.*[35] employed a genetic algorithm to empirically parameterize p$K_a$ values for carboxylic acids in proteins. Li *et al.*[36] presented an empirical method in a study of the molecular determinants of p$K_a$ values of Asp and Glu in the protein turkey ovomucoid third domain (OMTKY3). They took into account three types of p$K_a$ perturbation: hydrogen bonding, desolvation, and charge–charge interactions, in which hydrogen bonds were identified as the prime p$K_a$ determinant. However, the p$K_a$ values were determined by changing the pH within a range centered on the p$K_a$ value of the residue of interest, while hydrogen bonds were identified based on the X-ray or NMR structure determined at a given pH.

Here, we present a very simple model derived from statistical theory to predict protein p$K_a$ values. It is based on the premise that the p$K_a$ shift of ionizable residues in proteins is linearly correlated to the interaction between a given residue and its surrounding residues. Despite its relative simplicity, the model is quite able to reproduce experimental results. Its principal shortcoming is the dependence on the size of the data set. The current data set is too small to derive prediction coefficients for internal residues that tend to have highly shifted p$K_a$ values and are often catalytically important, and the coefficients derived from surface residues are not suitable for extending the application in internal residues.

## MATERIALS AND METHODS

### The p$K_a$ database

A large collection of high quality experimental p$K_a$ values is essential to the development and assessment of this new method. Researchers at the Edward Jenner Institute have been maintaining a protein p$K_a$ database named PPD (URL: http://www.jenner.ac.uk/PPD/), which is the largest and the most comprehensive collection of experimentally identified p$K_a$ values. Unfortunately, the PPD web site does not provide a downloadable database in flat text or other human-readable format. For convenience, we recompiled a subset of experimental p$K_a$ values derived from PPD.

By extracting p$K_a$ values from the PPD web site and manually validating the data using experimental 3D structures from PDB (URL: http://www.rcsb.org), we obtained a subset of experimentally measured p$K_a$ values containing 1122 p$K_a$ values belonging to 667 unique dissociable sites (as of July, 2006). To minimize the side-effect of potentially contradictory experimental data, we filtered our data set based on the following rules: (a) p$K_a$ values identified by NMR methods were accepted; (b) p$K_a$ values for a single site differing by more than 1.0 when determined by separate experiments or a standard deviation (SD) greater than 0.5 were filtered out; (c) p$K_a$ values with a sequential number recorded in PPD inconsistent with that of PDB due to, for example, mutants or lack of crystal data were also excluded. After filtration, we obtained a collection of 475 unique sites located in 73 protein chains in which each site has only one experimental p$K_a$ value assigned (if the site had multiple p$K_a$ values, the average value was assigned instead), listed in Table I. In the data set, the p$K_a$ values of 46 sites are unusual because of physical or chemical factors, such as salt bridges or disulfide bridges (Table II). To obtain reasonable parameters, these data were excluded from the fitting procedure and were predicted using the parameters obtained from the remaining sites. Figure 1 depicts the histogram of p$K_a$ values for the residues Asp, Glu, His, and Lys.

### Model for p$K_a$ calculation

The difference in p$K_a$ between an amino acid residue in a protein and that of the amino acid in solution can be expressed as:

$$\mathrm{p}K = \mathrm{p}K^{\mathrm{mod}} + \frac{1}{2.3RT}\Delta\Delta G \qquad (1)$$

In Eq. (1), $R$ is the gas constant and $T$ is the temperature (usually 298 K). The term $\Delta\Delta G$ is the difference in free energies of deprotonation in the model compound and in the protein. Because of the complex nature of $\Delta\Delta G$, the p$K_a$ shift can be viewed as consisting of two parts: (1) transfer from the model compound to the pro-

**Table I**
*Statistics for the 475 p$K_a$ Values*

| Residue | Number | Average | Min | Max | SD |
|---|---|---|---|---|---|
| Asp | 150 | 3.47 | 1.40 | 9.90 | 1.22 |
| Glu | 146 | 4.16 | 2.10 | 7.60 | 0.81 |
| His | 96 | 6.30 | 2.30 | 9.20 | 1.37 |
| Cys | 8 | 4.67 | 2.23 | 8.80 | 2.91 |
| Tyr | 28 | 9.90 | 2.09 | 12.50 | 2.62 |
| Lys | 47 | 10.04 | 2.21 | 12.05 | 1.73 |


DOI 10.1002/prot

**Table II**
*List of p$K_a$ Values with Large Deviation from the p$K_a$ Values of Model Compounds*

| PDB | Residue | | p$K_a$ | Note | Reference |
|---|---|---|---|---|---|
| 1A2P | A:73 | E | 2.10 | D93 forms a salt link with R69 and titrates at much lower pH values. | 6 |
| | A:93 | D | 2.00 | | |
| | A:101 | D | 2.00 | | |
| 1A91 | 61 | D | 7.0 | D61 is thought to protonate and deprotonate during each proton translocation cycle. This p$K_a$ was significantly higher. | 37 |
| 1CDC | A:41 | E | 6.7 | E41 is on the binding surface of rat CD2 with an unusually elevated p$K_a$, and the electrostatic interaction with the E29 side chain is a significant contributing influence. | 38 |
| 1D3K | A:45 | Y | 7.09 | The low p$K_a$ values of tyrosines are due to the electrostatic interactions with the neighboring groups. The low p$K_a$ of Y188 is due to the iron-binding ligand interactions with K206 in open-form and with K296 in the closed-form of the protein. | 39 |
| | A:85 | Y | 8.01 | | |
| | A:96 | Y | 7.06 | | |
| | A:188 | Y | 6.86 | | |
| 1HIC | 39 | C | 3.76 | | 40 |
| 1HPX | A:25 | D | 6.20 | The carboxyl of D25 (chain A) is protonated while that of D125 (chain B) is not protonated. The side chain of D25 is protonated in order to donate a hydrogen bond to carbonyl oxygen of KNI-272. | 41 |
| 1L63 | 31 | H | 9.10 | A salt bridge formed between the side chains of D70 and H31 contributes the perturbation of p$K_a$ values in the native state. | 42 |
| | 70 | D | 1.40 | | |
| 1LYS | A:66 | D | 2.00 | The p$K_a$ of D66 is inaccurate because of an insufficient number of chemical shift values obtained at low pH. | 43 |
| 1LZ3 | 66 | D | 2.00 | | |
| 1M8B | A:56 | C | 2.27 | | 44 |
| 1M8C | A:56 | C | 2.27 | | |
| 1MHI | A:19 | Y | 2.87 | Salt bridges are established between E (B13) and H (B10). | 45 |
| | B:5 | H | 3.75 | | |
| | B:10 | H | 2.66 | | |
| | B:13 | E | 2.45 | | |
| | B:16 | Y | 2.09 | | |
| 1MUT | 53 | E | 7.60 | E53 is functioning as a base catalyst in the active quaternary complex. | 46 |
| 1OMU | 31 | Y | 12.50 | C56 forms disulfide bridge with C24. The very high p$K_a$ of Y31 may be attributed to a short hydrogen bond with D27 and its low solvent accessible surface area. | 47 |
| | 56 | C | 2.50 | | |
| 1PNT | 72 | H | 9.20 | The high p$K_a$ is hypothesized to be caused by electrostatic interactions with nearby negatively charged groups (E23 and D42). | 48 |
| 1QH7 | A:60 | H | 4.01 | The low p$K_a$ of H60 provides a hint of an unusual microenvironment, despite the electrostatic influence of E167 which would be expected to elevate the p$K_a$. The low p$K_a$ of H162 reflects the buried hydrophobic environment of the imidazole ring within the enzyme. | 49 |
| | A:162 | H | 2.70 | | |
| 1RGG | A:79 | D | 7.37 | D79, which is buried but does not form hydrogen bonds, has the most elevated p$K_a$. | 10 |
| 1RNZ | 14 | D | 1.90 | | n/a |
| 1SBT | 17 | H | 3.00 | The four histidine residues are neutrally charged and do not titrate. | 50 |
| | 39 | H | 3.00 | | |
| | 67 | H | 3.00 | | |
| | 226 | H | 3.00 | | |
| 1SPU | A:383 | D | 9.70 | The high p$K_a$ is due to the fact that the deprotonation disrupts the hydrogen-bonding interaction with the pyridine nitrogen of the 2HP moiety. | 51 |
| 1SSO | 18 | K | 3.00 | K18 is involved in both a salt bridge and an H-bond. | 52 |
| 1TRS | 26 | D | 8.10 | In both the reduced and the oxidized states of human thioredoxin, | 53 |
| 1TRW | 26 | D | 9.90 | The stabilization of the protonated side chain of D26 is achieved via a hydrogen-bonding network involving the hydroxyl group of the neighboring S28. | |
| 1XNB | 83 | D | 2.00 | D83 is completely buried, forming a strong salt bridge with R136. | 54,55 |
| | 101 | D | 2.00 | D101 is located on the surface of the protein, stabilized in the deprotonated form by an extensive network of hydrogen bonds. | |
| | 149 | H | 2.30 | | |
| | 172 | E | 6.70 | H149 is completely buried within the hydrophobic core and is involved in an extensive network of hydrogen bonding interactions. E172 plays a role of the acid/base catalyst. | |
| 2OVO | 29 | K | 2.21 | The low p$K_a$ for the $\alpha$-carboxyl group of C56 is attributed to acidification by the disulfide group. | 56 |
| | 56 | C | 2.23 | | |
| 2RN2 | 102 | D | 2.00 | | 57 |
| | 148 | D | 2.00 | | |

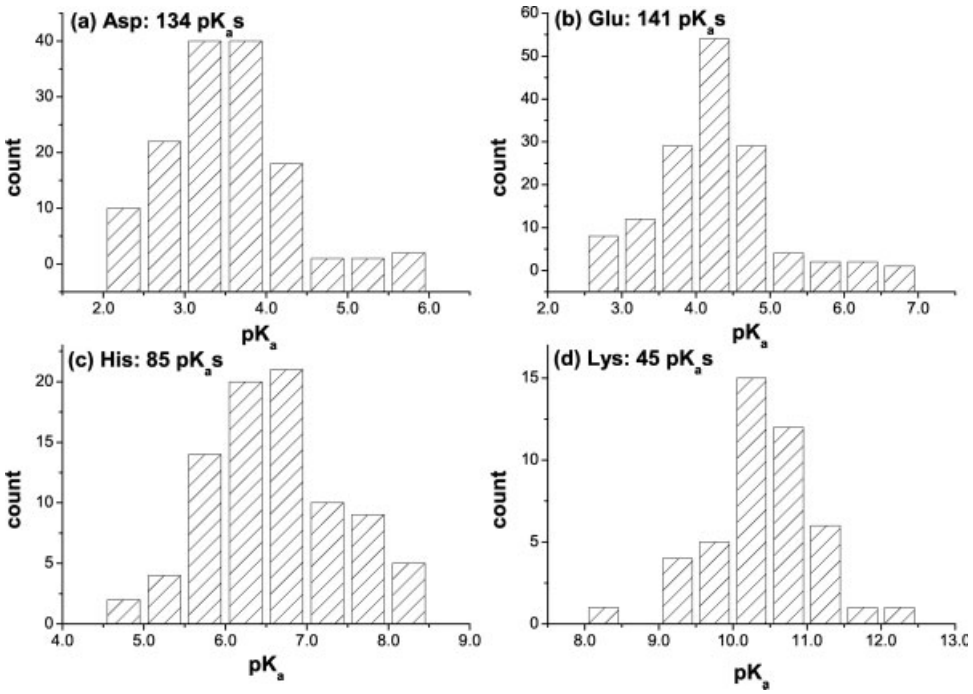


**Figure 1**

*Distributions of $pK_a$ values in proteins: (**a**) 134 Asp, (**b**) 141 Glu, (**c**) 85 His and (**d**) 45 Lys. Each column entry represents the number of $pK_a$ values, in 0.5 pH unit increments.*

tein neglecting interactions with other sites, giving the intrinsic $pK_a$ ($pK_a^{intr}$); (2) site–site interactions resulting in an additional shift in the $pK_a$.[17] Therefore, the term $\Delta\Delta G$ can be separated into a site–site interaction term ($\Delta G_{ss}$) and an intrinsic term ($\Delta G_{intr}$), and Eq. (1) can be rewritten as:

$$pK = pK^{mod} + \frac{1}{2.3RT}\Delta\Delta G = pK^{mod} + \frac{1}{2.3RT}(\Delta G_{ss} + \Delta G_{intr}) = pK^{intr} + \frac{1}{2.3RT}\Delta G_{ss} \quad (2)$$

Here, we adopt a simplified model for calculating $pK_a$. The main issues are as follows.

To avoid estimating $pK_a^{intr}$ in Eq. (2), we instead used $pK_a^0$ (the average $pK_a$ value found in the data set of the corresponding residues) as a reference state. Since the experimental $pK_a$ values display a normal distribution, the value of $pK_a^0$ differs from the intrinsic $pK_a^{intr}$ only by a constant.

The $pK_a$ value of a given residue $i$ can then be described as:

$$pK_i = pK_i^0 + \frac{1}{2.3RT}\Delta G_i \quad (3)$$

The model described here assumes that the interaction between the residue of interest and the surrounding residues is the dominant influence on $pK_a$, and that interactions with more distant residues can be neglected because of shielding by the neighboring residues. To simplify the calculation, we defined a sphere with a fixed radius centered at the $C\alpha$ atom of the residue of interest and calculated the $pK_a$ shift using only residues whose $C\alpha$ atom was located within the sphere.

NMR experiments to determine protein $pK_a$ values indicate that with a few exceptions, the pH dependency of chemical shifts of ionizable residues could be fitted by the Henderson-Hasselbach equation.[6–10] This indicates that the free energy change of the proton binding reaction is unchanged by variations in pH, although the residues undergo changes in charge state and the protein constantly experiences conformational fluctuations. We therefore assume that the $pK_a$ shift is a function of the amino acids in the sphere:

$$pK_i = pK_i^0 + f(n_1, n_2, \ldots, n_{20}) \quad (4)$$

Here $n_j$ ($j$ = 1–20) is the number of $j$th amino acid located in the sphere, and the function $f(n_1,n_2,\ldots,n_{20})$ is unknown.

The function $f(n_1,n_2,\ldots,n_{20})$ is obtained by a statistical approach. According to statistical principles, if $n_1$, $n_2,\ldots,n_{20}$ are independent random variables, their linear combination is a normal distribution. For simplification,


DOI 10.1002/prot

we assume that the p$K_a$ shift induced by residue–residue interaction is a linear function of the amino acids around the ionizable residue:

$$pK_i = pK_i^0 + \sum_{j=1}^{20} c_{ij} n_{ij} \quad (5)$$

Here $i$ ($i$ = 1–7) for the ionizable residues Asp, Glu, His, Cys, Tyr, Arg and Lys, and $j$ ($j$ = 1–20) for 20 amino acids with N-termini represented as an analog of Lys and C-termini as an analog of Asp. $C_{ij}$ denotes the coefficients for the residue–residue interaction term for residue $i$ with another residue $j$, and $n_{ij}$ denotes the number of residue $j$ around the residue $i$.

### Fitting procedure with multiple linear regression method

The coefficients $C_{ij}$ were determined from the data in the training set using the multiple linear regression method to minimize the sum of the square of deviations between the left and the right side of the equation.[58,59] In addition to a self-consistency validation, a sixfold cross validation was performed. The 73 protein chains were randomly placed into six groups and each group in turn was predicted using the coefficients obtained using the remaining groups as a training set. Since the data size for Cys and Tyr were too small to determine the coefficients $C_{ij}$, predictions of Cys and Tyr were performed using the coefficients of residue His.

## RESULTS

The model used here assumes that the interaction between the residue of interest and its close neighbors is the dominant factor in p$K_a$ shift, and interactions between this residue and more distant residues can be neglected because of shielding by adjacent residues. To define the first layer, a number of parameters such as the orientation and shielding of the residues should be used. Because of limitations on the size of the data set, an excessive number of parameters could result in over-fitting. To simplify the calculation, we defined a sphere of fixed volume centered at the C$\alpha$ atom of the ionizable residue.

**Table III**
*RMSD of Prediction with 20 Parameters for 405 p$K_a$ Values*

| Residue | Data size | Self-consistency validation | Sixfold cross validation |
|---|---|---|---|
| Asp | 134 | 0.557 | 0.728 |
| Glu | 141 | 0.507 | 0.659 |
| His | 85 | 0.705 | 1.111 |
| Lys | 45 | 0.425 | 1.845 |
| TOTAL | 405 | 0.562 | 0.982 |

**Table IV**
*RMSD of Prediction with 13 Parameters For 405 p$K_a$ Values*

| Residue | Data size | Self-consistency validation | Sixfold cross validation |
|---|---|---|---|
| Asp | 134 | 0.569 | 0.662 |
| Glu | 141 | 0.521 | 0.609 |
| His | 85 | 0.769 | 0.990 |
| Lys | 45 | 0.486 | 1.040 |
| TOTAL | 405 | 0.593 | 0.775 |

Residues whose C$\alpha$ atom was located within the sphere were considered in the p$K_a$ shift calculation. It was found that a sphere of radius 11 Å gave the minimum RMSD (data not show).

For each ionizable residue $i$ (Asp, Glu, His, Lys), 20 coefficients $C_{ij}$ were determined by fitting the data in the training set. The self-consistency and sixfold cross validation tests were performed in the prediction procedure, and the results are listed in Table III. The average RMSD between predicted and experimental values of all 405 sites for the self-consistency validation and the sixfold cross validation are 0.562 and 0.982, respectively. In common with previous prediction works,[60] the results demonstrate a gap in RMSD between the prediction performance of the self-consistency validation and the sixfold cross validation. The gap increases with decreasing size of the data set. The largest data set with 141 sites of residue Glu produced RMSD values of 0.507/0.659 for self-consistency/sixfold cross validation, while the smallest data set with 45 sites of residue Lys yielded RMSD values of 0.425/1.845.

The above results demonstrate that prediction using 20 parameters could result in over-fitting due to limitations of data set size. To avoid over-fitting, the parameters were reduced by classifying the 20 amino acids into 13 groups: His, Cys and Tyr were treated as His (with functional group —XH, X = O or N); Arg and Lys as Lys (with functional group: —$NH_3$); Asn and Gln as Asn (with functional group: —$CONH_2$); Ala, Gly, Leu, and Val as Ala (hydrophobic residues); and the rest remains ungrouped. Prediction using 13 parameters produced an obvious improvement in the sixfold cross validation as shown in Table IV. The RMSD of the sixfold cross validation for 405 p$K_a$ values decreased from 0.982 with 20 parameters to 0.775 with 13 parameters. The improvement in prediction for residue Lys was especially apparent, where the RMSD of the sixfold cross validation prediction decreased from 1.845 with 20 parameters to 1.0 with 13 parameters. Further attempts at reducing the amino acid alphabet or reclassification produced no obvious improvements in RMSD (data not shown). The correlation plots between experimental p$K_a$ values and predictions calculated with 13 parameters are presented in Figure 2. These plots show that most of the calculated

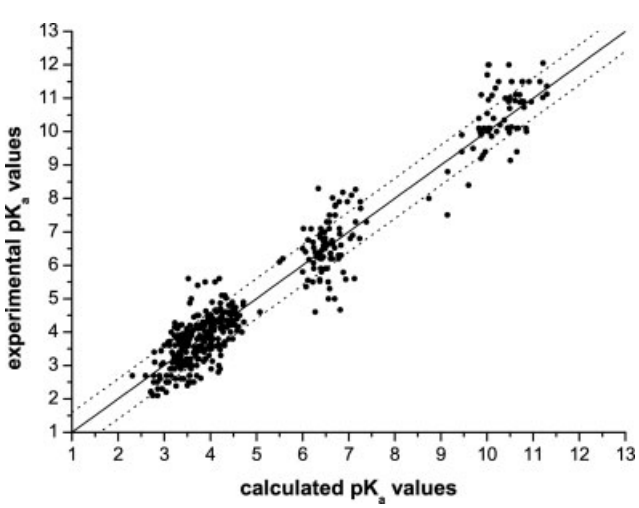


**Figure 2**

*Correlation plots between experimental and calculated pK$_a$ values for all 429 ionizable sites (134 Asp, 141 Glu, 85 His, 45 Lys, 3 Cys, and 21 Tyr). The solid line is the diagonal and the dotted lines represent 0.6 pH unit deviation from experimental values.*

values are within 0.6 pH units of the corresponding experimental values.

In the current data set, data for Cys and Tyr was insufficient for parameter fitting and data for residue Arg was not available. To extend the prediction program to these three residues, the coefficients for His were used for the prediction of Cys and Tyr, since they share the same functional group '—XH' (X = O or N). The RSMD of residues Cys and Tyr is ~1 pH unit (Table V). We therefore suggest that the coefficients for Lys could also be used for the prediction of residue Arg, considering the common functional group '—$NH_3$'.

Approximately 10% of the ionizable residues with highly shifted p$K_a$ values are located in an unusual environment. The real scientific challenge is to accurately predict these p$K_a$ values. Unfortunately, the present method based on statistical approach is not suitable to predict these p$K_a$ values due to the limitation of the data size. The average RMSD of prediction for these 46 sites is greater than 4 pH units, which indicates that these residues should be treated separately (the results are listed in Table VI and illustrated in Figure 3).

**Table V**

*RMSD of Prediction for Cys and Tyr with the Coefficients of His*

| Residue | Data size | 20 parameters | 13 parameters |
|---|---|---|---|
| Cys | 3 | 1.241 | 1.050 |
| Tyr | 21 | 1.080 | 1.067 |

**Table VI**

*RMSD of Prediction for 46 pK$_a$ Values Listed in Table II with 13 Parameters Obtained from Training Set of 405 pK$_a$ Values*

| Residue | Data size | RMSD |
|---|---|---|
| Asp | 16 | 3.252 |
| Glu | 5 | 2.112 |
| Lys | 2 | 7.723 |
| His | 11 | 3.442 |
| Cys | 5 | 6.666 |
| Tyr | 7 | 4.938 |
| Total | 46 | 4.258 |

## DISCUSSION

In this study, we propose a simple model for calculation of amino acid p$K_a$ values in proteins. Despite its simplicity, the model exhibits good prediction performance. Compared with other theoretical models and empirical prediction methods described in the literature, our model contains several significant differences.

In our model, p$K_a$ values derived from model compounds (p$K_a^{mod}$) are replaced by average values derived from a database (p$K_a^0$). The advantage of this replacement is that some specific parameters such as dielectric constants for the protein or the solvent are omitted. Because the experimentally determined p$K_a$ values show a normal distribution, the value of p$K_a^0$ differs from that of p$K_a^{mod}$ by a constant obtained automatically from the fitting procedure.

The outcome of p$K_a$ calculations based on macroscopic electrostatic models critically depends on the choice of protein dielectric constant, but defining this value is not straightforward. To establish the difference between vari-

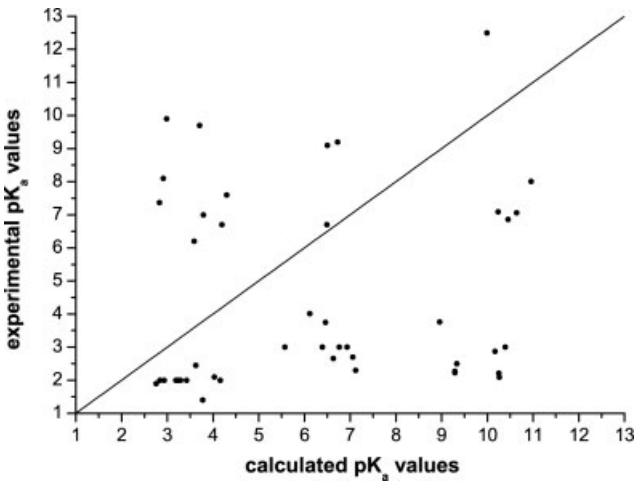


**Figure 3**

*Correlation plots between experimental and calculated pK$_a$ values for 46 ionizable sites which have large pK$_a$ shift as listed in Table II (16 Asp, 5 Glu, 11 His, 2 Lys, 5 Cys, and 7 Tyr). The solid line is the diagonal.*


DOI 10.1002/prot

ous macroscopic and semimacroscopic models and to illustrate the nature of the protein dielectric constants, Schutz and Warshel[19] proposed a discriminative benchmark that mainly included residues whose p$K_a$ values were shifted significantly from their values in water, and they asserted that the optimal dielectric constant for self-energies is not the optimal dielectric constant for charge–charge interactions. The protein dielectric constant is not a universal constant but simply a parameter that depends on the model used. In our study, the $\Delta G$ term in the p$K_a$ calculation appears in a very simple form with only 20 (or 13 in a reduced alphabet) residue-type parameters. Because proteins are made of amino acid building blocks, residue–residue interactions include not only electrostatic interactions but also other types such as van der Waals interactions, which are usually ignored by models described in the literature. Furthermore, in this model, the $\Delta G$ term of the p$K_a$ shift is proportional to the number of amino acids surrounding the residue of interest, since buried residues are surrounded by more neighbors than residues exposed to solvent. This is consistent with previous studies asserting that a relationship exists between solvent accessibility surface area (ASA) and p$K_a$ values: p$K_a$ values tend to decrease with increasing ASA.[5]

During NMR experiments to determine p$K_a$ values, there are changes in the charge states of the residue of interest and surrounding residues, and the protein is constantly undergoing conformational fluctuations in response to changing solution pH. However, techniques for including conformational flexibility in titration calculations are still under development. The methods applied so far have several shortcomings; for instance, the chosen conformational ensemble does not necessarily adequately represent the protein conformational space over the investigated pH range, and no conformational variation of the protein backbone is currently allowed during the computational titration.[17] The NMR experiments for determining protein p$K_a$ values show that with a few exceptions, the pH dependency of the chemical shifts of the ionizable residues could be fitted by the Henderson-Hasselbach equation.[6–10] This suggests that the free energy change of the proton binding reaction is unchanged during the change in pH. It is most likely that the protein structure does not sample all the conformations, rather only the ensemble whose proton binding free energy change can be compensated by conformational change. Despite the fact that the model presented here ignores both conformational changes and changes of the residue charge state, it performs quite well in predicting p$K_a$ values.

Despite its relative simplicity, the model is quite capable of reproducing experimental results. The principal shortcoming of this statistical model is the dependence on the size of the data set. In order to reduce fitting parameters, the model neglects the orientation and shielding of residues surrounding the amino acid of interest, and this omission could introduce errors into the prediction. Furthermore, the current data set is too small to derive prediction coefficients for Cys and Tyr as well as residues with great p$K_a$ shift. An increase in the number of experimentally determined p$K_a$ values would overcome this shortcoming and further enhance the accuracy of our model.

## REFERENCES


1. Perez-Jimenez R, Godoy-Ruiz R, Ibarra-Molero B, Sanchez-Ruiz JM. The efficiency of different salts to screen charge interactions in proteins: a Hofmeister effect? Biophys J 2004;86:2414–2429.
2. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein–protein interactions. Curr Opin Struct Biol 2000;10:153–159.
3. Elcock AH. Realistic modeling of the denatured states of proteins allows accurate calculations of the pH dependence of protein stability. J Mol Biol 1999;294:1051–1062.
4. Warshel A, Sharma PK, Kato M, Parson WW. Modeling electrostatic effects in proteins. Biochim Biophys Acta 2006;1764:1647–1676.
5. Forsyth WR, Antosiewicz JM, Robertson AD. Empirical relationships between protein structure and carboxyl pKa values in proteins. Proteins 2002;48:388–403.
6. Oliveberg M, Arcus VL, Fersht AR. pKa values of carboxyl groups in the native and denatured states of barnase: the pKa values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry 1995;34:9424–9433.
7. Kesvatera T, Jonsson B, Thulin E, Linse S. Measurement and modelling of sequence-specific pKa values of lysine residues in calbindin D9k. J Mol Biol 1996;259:828–839.
8. Spitzner N, Lohr F, Pfeiffer S, Koumanov A, Karshikoff A, Ruterjans H. Ionization properties of titratable groups in ribonuclease T1. I. pKa values in the native state determined by two-dimensional heteronuclear NMR spectroscopy. Eur Biophys J 2001;30:186–197.
9. Hatano K, Kojima M, Tanokura M, Takahashi K. Nuclear magnetic resonance studies on the pKa values and interaction of ionizable groups in bromelain inhibitor VI from pineapple stem. Biol Chem 2003;384:93–104.
10. Laurents DV, Huyghues-Despointes BM, Bruix M, Thurlkill RL, Schell D, Newsom S, Grimsley GR, Shaw KL, Trevino S, Rico M, Briggs JM, Antosiewicz JM, Scholtz JM, Pace CN. 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.
11. Warshel A, Sussman F, King G. Free energy of charges in solvated proteins: microscopic calculations using a reversible charging process. Biochemistry 1986;25:8368–8372.
12. Warshel A, Russell ST. Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys 1984;17:283–422.
13. Sham YY, Chu ZT, Warshel A. Consistent Calculations of pKa's of ionizable residues in proteins: semi-microscopic and microscopic approaches. J Phys Chem B 1997;101:4458–4472.
14. Tanford C, Kirkwood JG. Theory of protein titration curves. I. General equations for impenetrable spheres. J Am Chem Soc 1957;79:5333–5339.
15. Warshel A, Russell ST, Churg AK. Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. Proc Natl Acad Sci USA 1984;81:4785–4789.
16. Sharp KA, Honig B. Electrostatic interactions in macromolecules: theory and applications. Annu Rev Biophys Biophys Chem 1990;19:301–332.
17. Juffer AH. Theoretical calculations of acid-dissociation constants of proteins. Biochem Cell Biol 1998;76:198–209.
18. Warshel A, Papazyan A. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. Curr Opin Struct Biol 1998;8:211–217.
19. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.

20. Warshel A. Calculations of enzymatic reactions: calculations of pKa, proton transfer reactions, and general acid catalysis reactions in enzymes. Biochemistry 1981;20:3167–3177.
21. Russell ST, Warshel A. Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. J Mol Biol 1985;185:389–404.
22. King G, Lee FS, Warshel A. Microscopic simulations of macroscopic dielectric constants of solvated proteins. J Chem Phys 1991;95: 4366–4377.
23. Zhou HX, Vijayakumar M. Modeling of protein conformational fluctuations in pKa predictions. J Mol Biol 1997;267:1002–1011.
24. Ullmann GM, Knapp EW. Electrostatic models for computing protonation and redox equilibria in proteins. Eur Biophys J 1999;28:533–551.
25. Koumanov A, Ruterjans H, Karshikoff A. Continuum electrostatic analysis of irregular ionization and proton allocation in proteins. Proteins 2002;46:85–96.
26. Mehler EL, Fuxreiter M, Simon I, Garcia-Moreno EB. The role of hydrophobic microenvironments in modulating pKa shifts in proteins. Proteins 2002;48:283–292.
27. Dong F, Vijayakumar M, Zhou HX. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. Biophys J 2003;85:49–60.
28. Kuhn B, Kollman PA, Stahl M. Prediction of pKa shifts in proteins using a combination of molecular mechanical and continuum solvent calculations. J Comput Chem 2004;25:1865–1872.
29. Lee MS, Salsbury FR, Brooks CL. Constant-pH molecular dynamics using continuous titration coordinates. Proteins 2004;56:738–752.
30. Simonson T, Carlsson J, Case DA. Proton binding to proteins: pKa calculations with explicit and implicit solvent models. J Am Chem Soc 2004;126:4167–4180.
31. Warwicker J. Improved pKa calculations through flexibility based sampling of a water-dominated interaction scheme. Protein Sci 2004;13:2793–2805.
32. Riccardi D, Schaefer P, Cui Q. pKa calculations in solution and proteins with QM/MM free energy perturbation simulations: a quantitative test of QM/MM protocols. J Phys Chem B 2005;109:17715–17733.
33. Ohno K, Sakurai M. Linear-scaling molecular orbital calculations for the pKa values of ionizable residues in proteins. J Comput Chem 2006;27:906–916.
34. Wisz MS, Hellinga HW. An empirical model for electrostatic interactions in proteins incorporating multiple geometry-dependent dielectric constants. Proteins 2003;51:360–377.
35. Godoy-Ruiz R, Perez-Jimenez R, Garcia-Mira MM, Plaza del Pino IM, Sanchez-Ruiz JM. Empirical parametrization of pK values for carboxylic acids in proteins using a genetic algorithm. Biophys Chem 2005;115:263–266.
36. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins 2005;61:704–721.
37. Assadi-Porter FM, Fillingame RH. Proton-translocating carboxyl of subunit c of F1Fo H(+)-ATP synthase: the unique environment suggested by the pKa determined by 1H NMR. Biochemistry 1995;34:16186–16193.
38. Chen HA, Pfuhl M, McAlister MS, Driscoll PC. Determination of pKa values of carboxyl groups in the N-terminal domain of rat CD2: anomalous pKa of a glutamate on the ligand-binding surface. Biochemistry 2000;39:6814–6824.
39. Sun X, Sun H, Ge R, Richter M, Woodworth RC, Mason AB, He QY The low pKa value of iron-binding ligand Tyr188 and its implication in iron release and anion binding of human transferrin. FEBS Lett 2004;573:181–185.
40. Szyperski T, Antuch W, Schick M, Betz A, Stone SR, Wuthrich K. Transient hydrogen bonds identified on the surface of the NMR solution structure of Hirudin. Biochemistry 1994;33:9303–9310.
41. Wang YX, Freedberg DI, Yamazaki T, Wingfield PT, Stahl SJ, Kaufman JD, Kiso Y, Torchia DA. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. Biochemistry 1996;35:9945–9950.
42. Anderson DE, Becktel WJ, Dahlquist FW. pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry 1990;29:2403–2408.
43. Bartik K, Redfield C, Dobson CM. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional 1H NMR. Biophys J 1994;66:1180–1184.
44. Song J, Laskowski M, Qasim MA, Markley JL. Two conformational states of Turkey ovomucoid third domain at low pH: three-dimensional structures, internal dynamics, and interconversion kinetics and thermodynamics. Biochemistry 2003;42:6380–6391.
45. Sorensen MD, Led JJ. Structural details of Asp(B9) human insulin at low pH from two-dimensional NMR titration studies. Biochemistry 1994;33:13727–13733.
46. Harris TK, Wu G, Massiah MA, Mildvan AS. Mutational, kinetic, and NMR studies of the roles of conserved glutamate residues and of lysine-39 in the mechanism of the MutT pyrophosphohydrolase. Biochemistry 2000;39:1655–1674.
47. 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.
48. Tishmack PA, Bashford D, Harms E, Van Etten RL. Use of 1H NMR spectroscopy and computer simulations to analyze histidine pKa changes in a protein tyrosine phosphatase: experimental and theoretical determination of electrostatic properties in a small protein. Biochemistry 1997;36:11984–11994.
49. 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.
50. Day RM, Thalhauser CJ, Sudmeier JL, Vincent MP, Torchilin EV, Sanford DG, Bachovchin CW, Bachovchin WW. Tautomerism, acid-base equilibria, and H-bonding of the six histidines in subtilisin BPN' by NMR. Protein Sci 2003;12:794–810.
51. Mure M, Brown DE, Saysell C, Rogers MS, Wilmot CM, Kurtis CR, McPherson MJ, Phillips SE, Knowles PF, Dooley DM. Role of the interactions between the active site base and the substrate Schiff base in amine oxidase catalysis. Evidence from structural and spectroscopic studies of the 2-hydrazinopyridine adduct of *Escherichia coli* amine oxidase. Biochemistry 2005;44:1568–1582.
52. Consonni R, Arosio I, Belloni B, Fogolari F, Fusi P, Shehi E, Zetta L. Investigations of Sso7d catalytic residues by NMR titration shifts and electrostatic calculations. Biochemistry 2003;42:1421–1429.
53. 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 κB. Biochemistry 1996;35:7–13.
54. Joshi MD, Hedberg A, McIntosh LP. Complete measurement of the pKa values of the carboxyl and imidazole groups in *Bacillus circulans* xylanase. Protein Sci 1997;6:2667–2670.
55. Joshi MD, Sidhu G, Pot I, Brayer GD, Withers SG, McIntosh LP. Hydrogen bonding and catalysis: a novel explanation for how a single amino acid substitution can change the pH optimum of a glycosidase. J Mol Biol 2000;299:255–279.
56. Schaller W, Robertson AD. pH, ionic strength, and temperature dependences of ionization equilibria for the carboxyl groups in turkey ovomucoid third domain. Biochemistry 1995;34:4714–4723.
57. 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.
58. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
59. Pan XM. Multiple linear regression for protein secondary structure prediction. Proteins 2001;43:256–259.
60. Wang ZX, Yuan Z. How good is prediction of protein structural class by the component-coupled method? Proteins 2000;38:165–175.

 DOI 10.1002/prot