Protein Engineering, Design & Selection vol. 23 no. 1 pp. 35–42, 2010
Published online November 18, 2009 doi:10.1093/protein/gzp067

A fast and accurate method for predicting pK_a of residues in proteins

Ri-Bo Huang[1,2,3], Qi-Shi Du[1,2,4,5], Cheng-Hua Wang[3], Si-Ming Liao[2] and Kuo-Chen Chou[4]

[1]Guangxi Academy of Sciences, 98 Daling Road, Nanning, Guangxi 530004, People's Republic of China, [2]College of Life Science and Technique, Guangxi University, Nanning, Guangxi 530004, People's Republic of China, [3]College of Life Science and Pharmaceutical Engineering, Nanjing University of Technology, Nanjing, Jiangsu 210009, People's Republic of China and [4]Gordon Life Science Institute, San Diego, CA 92130, USA

[5]To whom correspondence should be addressed. Tel: (086) 771-250-3931; Fax: (086) 771-250-3908; E-mail: qishi_du@yahoo.com.cn

Received May 18, 2009; revised August 19, 2009; accepted October 20, 2009

Edited by P. Balaram

Predicting the pH-activities of residues in proteins is an important problem in enzyme engineering and protein design. A novel predictor called 'Pred-pK_a' was developed based on the physicochemical properties of amino acids and protein 3D structure. The Pred-pK_a approach considers the influence of all other residues of the protein to predict the pK_a value of an ionizable residue. An empirical equation was formulated, in which the pK_a value was a distance-dependent function of physicochemical parameters of 20 amino acid types, describing their electrostatic and van der Waals interaction, as well as the effects of hydrogen bonds and solvation. Two sets of coefficients, $\{a_\alpha\}$ and $\{b_l\}$, were used in the predictor: $\{a_\alpha\}$ is the weight factors of 20 amino acid types and $\{b_l\}$ is the weight factors of physicochemical properties of amino acids. An iterative double least square procedure was proposed to solve the two sets of weight factors alternately and iteratively in a training set. The two coefficient sets $\{a_\alpha\}$ and $\{b_l\}$ thus obtained were used to predict the pK_a values of residues in a protein. The average predictive error is ± 0.6 pH in less than a minute in common personal computer.

Keywords: acid ionic constant/drug design/pH-activity/protein engineering/residue mutations

Introduction

The pK_a values (ionization constants) of the residues in active sites of an enzyme are of importance to the functionality of the catalytic mechanism of the enzyme, which govern the pH dependence of protein stability and enzymatic activity. Usually, an enzyme catalytic reaction is initiated by the transfer of a proton from a protein residue (the proton donor) to the substrate. One of the following steps in the enzymatic reaction mechanism is normally a nucleophilic attack on a substrate atom or the stabilization of a positively charged intermediate. The pK_a values of the proton donor and the catalytic nucleophilic residue are therefore limiting the pH range at which the enzyme can function. A large number of mutagenesis experiments have been carried out in the last three decades to re-engineer the pH-activity and pH-stability profile of enzymes and proteins. Therefore, the prediction of residue pK_a values in enzymes is of major importance for the biological function and industrial application of enzymes and proteins.

Currently, the pK_a prediction methods fall into two main categories: theoretical calculation and empirical prediction. In the theoretical approach, the most popular pK_a prediction methods (Bashford and Karplus, 1990; Yang *et al*., 1993; Antosiewicz *et al*., 1994, 1996; Mehler and Guarnieri, 1999; Ullmann and Knapp, 1999) are based on electrostatic continuum models that numerically solve the linearized Poisson–Boltzmann equation (LPBE),

$$\nabla\varepsilon(r)\nabla\phi(r) - \kappa^2(r)\varepsilon(r)\phi(r) = -4\pi\rho(r) \qquad (1)$$

where ε is the dielectric constant, r the position vector, $\varphi(r)$ the electrostatic potential, $\rho(r)$ the charge distribution and κ a parameter that represents the effect of mobile ions in solution. In these methods, the protein is described 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. The Poisson–Boltzmann equation is solved numerically. Currently three programs are available for theoretical calculation of pK_a. They are MCCE (Alexov and Gunner, 1997, 1999), MEAD (Bashford, 1997) and UHBD (Madura *et al*., 1995).

In the empirical prediction method, the pK_a value of an ionizable group in a protein is predicted by applying an environmental perturbation ΔpK_a to the model value of a residue in protein,

$$pK_a = pK_{model} + \Delta pK_a \qquad (2)$$

The environmental perturbation ΔpK_a is supported by the sum of several physical and chemical contributions, including hydrogen bond (HB), desolvation (Des) and charge–charge (q–q) interaction,

$$\begin{aligned}\Delta pK_a &= \Delta pK_{HB} + \Delta pK_{Des} + \Delta pK_{q-q} \\ &= C_{HB}f_{HB}(r)N_{HB} + C_{Des}f_{Des}(r)N_{Des} + C_{q-q}f_{q-q}(r)N_{q-q}\end{aligned} \qquad (3)$$

where N_{HB}, N_{Des} and N_{q-q} are the physicochemical parameters of amino acids; $f_{HB}(r)$, $f_{Des}(r)$ and $f_{q-q}(r)$ the distance-dependent functions and C_{HB}, C_{Des} and C_{q-q} the weight factors of physicochemical parameters. Once the weight factors are determined in a training set, Equation (3) can be used to predict the pK_a of residues in other proteins. The available free software for empirical prediction is PROPKA (Li *et al*., 2005), which was developed by Jensen's group.

© The Author 2009. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oxfordjournals.org

The theoretical approach for pK_a calculations seems more advanced than the empirical approach. However, the current techniques used in solving the Poisson–Boltzmann equation contain some approximations, which cannot provide satisfied prediction for pK_a values. A careful comparison was performed by Davies *et al*. (2006) between empirical method PROPKA (Li *et al*., 2005) and three theoretical methods MCCE (Alexov and Gunner, 1997, 1999), MEAD (Bashford, 1997) and UHBD (Madura *et al*., 1995). Their conclusion was that the empirical method PROPKA made better pK_a predictions (Davies *et al*., 2006) than other three theoretical methods did. However, these models (Alexov and Gunner, 1997, 1999; Bashford, 1997; Harris and Turner, 2002; Li *et al*., 2005) usually have a root-mean-square deviation (RMSD) around 1 pH unit from experimental values, which is still too big for practical applications.

In this article, we present an improved empirical pK_a prediction method, Pred-pK_a, based on physicochemical properties of amino acids and 3D structure of proteins. We show that this empirical method is able to provide more accurate pK_a predictions than other available theoretical and empirical methods do. The empirical pK_a prediction method Pred-pK_a can predict most protein pK_a values and work in common PC computers in a matter of seconds.

Method

Theoretical model of Pred-pK_a

The Pred-pK_a method uses the same basic equation [Equation (2)] as used in the PROPKA (Li *et al*., 2005) program. However, Pred-pK_a does not use the model values of residues, but the average values, which are arithmetical mean of experimental pK_a values of residues in a training set,

$$pK_a^k = pK_{Aver} + \Delta pK_k \quad (k = 1, 2, \ldots, K) \tag{4}$$

The environmental perturbation ΔpK_k to the pK_a^k of residue k in a protein is supposed by the contributions of 20 types of amino acids,

$$\Delta pK_k = \sum_{\alpha=1}^{20} a_\alpha G_\alpha = \sum_{\alpha=1}^{20} a_\alpha \left[\sum_{i=1}^{N_\alpha} g_{\alpha,i}(\mathbf{r}_{k,i}) \right] \quad (k = 1, 2, \ldots, K) \tag{5}$$

where G_α is the contribution of α-type amino acid, $g_{\alpha,i}(\mathbf{r}_{k,i})$ the contribution of ith residue of α-type amino acid, $\mathbf{r}_{k,i}$ the distance and orientation between the residue k and the ith residue, N_α the residue number of α-type amino acid in a protein, a_α the weight factor of α-type amino acid. Here, we use the index k for the residues whose pK_a to be predicted; index α for the types of amino acids; index i is the serial number of residues in a protein. The contribution term $g_{\alpha,i}(\mathbf{r}_{k,i})$ of the ith residue of α-type amino acid is evaluated by its physicochemical parameters $u_{\alpha,i,l}$ and a distance-dependent function $f_l(\mathbf{r}_{k,i})$,

$$g_{\alpha,i}(\mathbf{r}_{k,i}) = \sum_{l=1}^{L} b_l f_l(\mathbf{r}_{k,i}) u_{\alpha,i,l} \tag{6}$$

where l is the index for physicochemical properties, $u_{\alpha,i,l}$ the lth physicochemical parameter of residue i of α-type amino acids, b_l the weight factor of the lth physicochemical parameter, L the number of physicochemical parameters used in the pK_a^k calculation. The physicochemical parameters $u_{\alpha,i,l}$ describe the electrostatic interaction, van der Waals interaction, solvation and desolvation interaction and hydrogen-bond interaction. Inserting Equation (6) into Equation (5) and reordering the equation, we get the following foundational equation of Pred-pK_a,

$$\sum_{\alpha=1}^{20} \left\{ a_\alpha \sum_{l=1}^{L} b_l \left[\sum_{i=1}^{N_\alpha} f_l(\mathbf{r}_{k,i}) u_{\alpha,i,l} \right] \right\} = \Delta pK_k \quad (k = 1, 2, \ldots, K) \tag{7}$$

In Equation (7), the terms in square bracket can be calculated based on the physicochemical parameters of amino acids and 3D structure of protein. The detailed calculation method will be introduced in next section. These terms in square bracket are simply denoted by $p_{k,\alpha,l}$,

$$p_{k,\alpha,l} = \sum_{i=1}^{N_\alpha} f_l(\mathbf{r}_{k,i}) u_{\alpha,i,l} \tag{8}$$

The notation $p_{k,\alpha,l}$ can be understood as the contribution of lth physicochemical property of the ith residues of α-type amino acid to ΔpK_k of residue k. After placing Equation (8) into the Equation (7), the foundational equation of Pred-pK_a is reformed as follows,

$$\sum_{\alpha=1}^{20} a_\alpha \left(\sum_{l=1}^{L} b_l p_{k,\alpha,l} \right) = \Delta pK_k \quad (k = 1, 2, \ldots, K) \tag{9}$$

The residue k, whose pK_a^k to be predicted, includes Asp, Glu, His, Cys, Tyr, Lys, Arg and other ionizable residues in proteins.

Calculations of terms $p_{k,\alpha,l}$

The contribution of physicochemical property $p_{k,\alpha,l}$ of amino acid type α and physical parameter l to ionizable residue k is computed using a set of physicochemical parameters of amino acids and 3D structure of protein according to Equation (8). These types of physicochemical properties of amino acids have been very useful in empirical quantitative structure–properties relationship studies in proteins. In these studies, we can use the properties of amino acids either directly or as inputs for the model. Indirect use involves calculating properties of the full protein and using them later as inputs. For instance, González-Díaz *et al*. have introduced a new method to calculate hydrophobic, van der Waals, HINT and electrostatic properties of proteins for structure-property studies based on amino acid properties (González-Díaz and Uriarte, 2005; Gonzalez-Diaz *et al*., 2005, 2007a,b; Concu *et al*., 2009). These authors have also reviewed many different approaches to derive predictors from amino acid and/or full protein parameters (González-Díaz *et al*., 2007, 2008a,b).

The distance and orientation functions $f_l(\mathbf{r}_{k,i})$ in Equation (8) are designed based on the natures of physicochemical parameters. The general form of distance functions $f_l(\mathbf{r}_{k,i})$ is

as follows,

$$f_l(\mathbf{r}_{k,i}) = R_l(r_{k,i})Y_l(\theta_{k,i}) \quad (10)$$

The radial function $R_l(r_{k,i})$ is a distance decaying function,

$$R_l(r_{k,i}) = \frac{1}{|r_k - r_i|^{n_l}} \quad (11)$$

where r_k and r_i are the mass centers of residue k and i, respectively. The exponents n_l of physicochemical properties are different based on the physical nature of the properties. For charge–charge interaction, the exponent n_l is 1 according to the Coulomb law, for dipole–dipole interaction it is 2, for attractive van der Waals interaction it is 6 and for hydrogen-bond interaction it is 4. For most physicochemical properties, the directional function $Y_l(\theta_{k,i})$ is constant 1. However, for hydrogen-bond interaction, it is a cosine function,

$$Y_l(\theta_{k,i}) = -\cos\theta_{k,i} \quad (12)$$

where $\theta_{k,i}$ is the bond angle of the hydrogen bond between residues k and i.

Iterative double least square

In the training calculations, the simultaneous linear equation (9) has two sets of unknown variables: $\{a_\alpha\}$ are weight factors of the 20 amino acids and $\{b_l\}$ the weight factors of the physicochemical properties. An iterative double least square (IDLS) technique was developed to determine the values of the two sets of weight factors $\{a_\alpha\}$ and $\{b_l\}$ alternately and iteratively. In Equation (9), the terms $p_{k,\alpha,l}$ form a 3D matrix $\mathbf{P}_{KML}$. Here, K is the number of samples of pK$_a$ in a training set, $M = 20$ is the number of amino acid types and L is the number of physicochemical properties. By using a set of initial values of weight factors $\{a_\alpha^{(0)} = 1, \alpha = 1, 2, \ldots, M\}$, the 3D data matrix $\mathbf{P}_{KML}$ is reduced to a 2D data matrix $\mathbf{D}^{(1)}_{K\times L}$ with the elements given by

$$d^{(1)}_{k,l} = \sum_{\alpha=1}^{M} a^{(0)}_\alpha p_{k,\alpha,l} \quad (13)$$

Thus, the 3D simultaneous linear equation set [Equation (9)] is reduced to a set of 2D equations; i.e.

$$\sum_{l=1}^{L} b^{(1)}_l d^{(1)}_{k,l} = \Delta\mathrm{pK}_k \quad (k = 1, 2, \ldots, K) \quad (14)$$

The above equation set can be solved by using the least square (LS) approach, yielding the first solutions of the weight factors $\{b^{(1)}_l\}$. Then the values of $\{b^{(1)}_l\}$ are used to reduce the 3D data matrix $\mathbf{P}_{KML}$ to a 2D data matrix $\mathbf{T}^{(1)}_{K\times L}$ with the elements given by

$$t^{(1)}_{k,\alpha} = \sum_{l=1}^{L} b^{(1)}_l p_{k,\alpha,l} \quad (15)$$

Similarly, the 3D simultaneous linear equation set [Equation (9)] is reduced to a 2D equation set by Equation (15), as given by

$$\sum_{\alpha=1}^{M} a^{(1)}_\alpha t^{(1)}_{k,\alpha} = \Delta\mathrm{pK}_k \quad (k = 1, 2, \ldots, K) \quad (16)$$

The above linear equation set can be solved by using the LS approach, leading to the solution of weight factors $\{a^{(0)}_\alpha\}$. Then the values of $\{a^{(0)}_\alpha\}$ are used for the new solutions of the weight factors $\{b^{(2)}_l\}$ of the physicochemical properties.

Converge and prediction

The above procedure is performed iteratively for n steps, i.e. until reaching the converged solutions as denoted by $\{a^{(n)}_\alpha\}$ and $\{b^{(n)}_l\}$. The convergence criterion for the iterative procedure is given by the following equation

$$\left|Q^{(n+1)} - Q^{(n)}\right| = \left|\sqrt{\frac{1}{K}\sum_{k=1}^{K}(\Delta\mathrm{pK}^{\mathrm{expt}}_k - \Delta\mathrm{pK}^{(n+1)}_k)^2} - \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\Delta\mathrm{pK}^{\mathrm{expt}}_k - \Delta\mathrm{pK}^{(n)}_k)^2}\right| \le \varepsilon(10^{-6}) \quad (17)$$

where $Q^{(n)}$ represents the square root of the summation of squared differences between the experimental pK$_a$ values and the predicted values in the nth step, and $Q^{(n+1)}$ that in the $(n+1)$th step. Now, the values of $\{a^{(n)}_\alpha\}$ and $\{b^{(n)}_l\}$ can be used to predict the ΔpK$_j$ of the jth query residue through the following equation:

$$\Delta\mathrm{pK}_j = \sum_{\alpha=1}^{M} a^{(n)}_\alpha \left(\sum_{l=1}^{L} b^{(n)}_l p_{j,\alpha,l}\right) \quad (18)$$

In the IDLS procedure, we need a set of initial values of weight factors $\{a^{(0)}_\alpha\}$ as the starting point of the iteration procedure. A reasonable guess for the initial values is $\{a^{(0)}_\alpha = 1, \alpha = 1, 2, \ldots, 20\}$ for all 20 amino acid types. This implies that all amino acids are equally important at the beginning of the iteration. The mathematical procedure of IDLS is schematically illustrated in flowchart Fig. 1.

Results

The first step of the Pred-pK$_a$ procedure is to build a training set containing reliable experimental pK$_a$ data of residues in an extensive scale. The training set is built using PDB files of 62 proteins, which are collected from 50 references (Meadows *et al*., 1969; Ruterjans and Witzel, 1969; Czerlinski and Dar, 1971; Cohen *et al*., 1973; Snyder *et al*., 1975; Brown *et al*., 1976; Kuramitsu and Hamaguchi, 1979, 1980; Walters, and Allerhand, 1980; Inagaki *et al*., 1981; March *et al*., 1982; Matthew *et al*., 1985; McNutt *et al*., 1990; Atkins *et al*., 1993; Loewenthal *et al*., 1993; Oda *et al*., 1993, 1994; Bartik *et al*., 1994; Sorensen and Led, 1994; Szyperski *et al*., 1994; Assadi-Porter and Fillingame, 1995; Oliveberg *et al*., 1995; Schaller and Robertson, 1995; Antosiewicz *et al*., 1996; Chiang *et al*., 1996; Kesvatera *et al*., 1996, 1999; Qin *et al*., 1996; Joshi *et al*., 1997; Khare *et al*., 1997; Forsyth *et al*., 1998; Gooley *et al*., 1998;

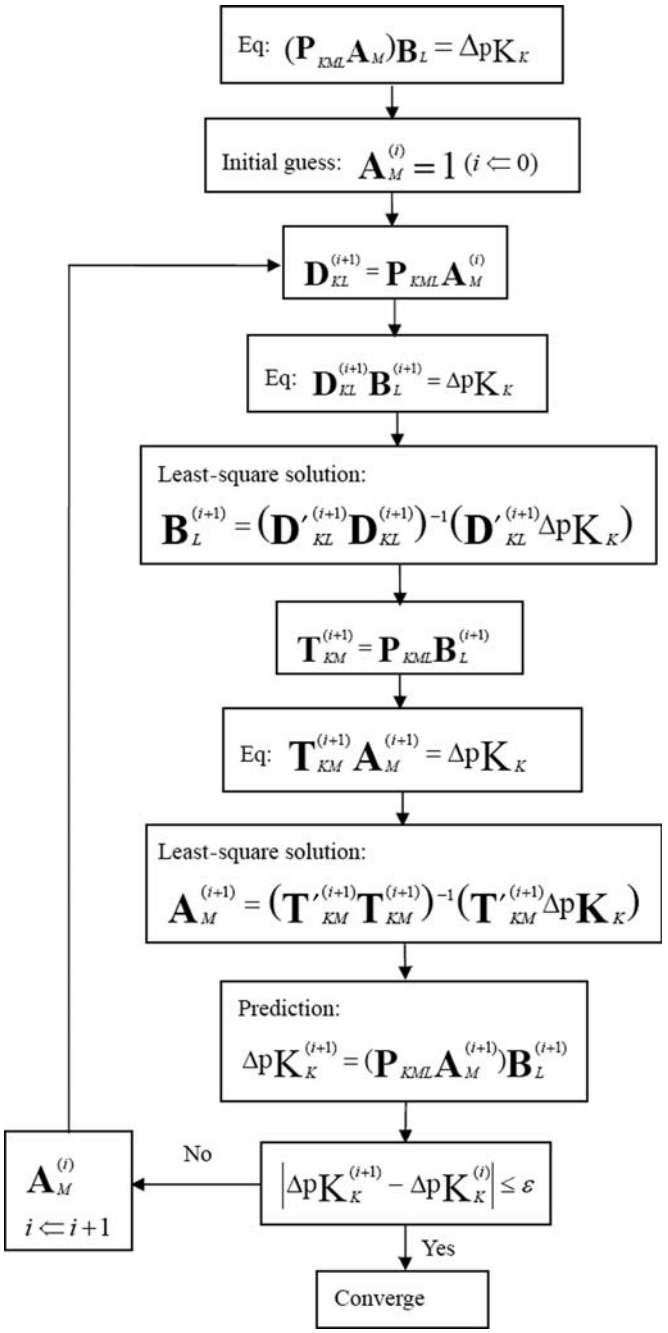

Fig. 1. The IDLS procedure for the solution of the 3D linear equations for two sets of coefficients $\mathbf{A}_M^{(n)}$ and $\mathbf{B}_L^{(n)}$. In the IDLS procedure K is the number of samples of pK_a in a training set, $M = 20$ is the number of amino acid types, and L is the number of parameters. $\mathbf{P}_{KML}$ is the three dimensional data matrix, $\mathbf{A}_M$ and $\mathbf{B}_L$ is the vector representing the coefficient set for 20 types of amino acids and for physicochemical parameters, respectively.

Perez-Canadillas *et al*., 1998; Gustafsson *et al*., 1999; Mossner *et al*., 2000; Bombarda *et al*., 2001, 2002; Gerratana *et al*., 2001; Hubatsch and Mannervik, 2001; Reyes-Vivas *et al*., 2001; Soulere *et al*., 2001; Georgescu *et al*., 2002; Consonni *et al*., 2003; Ibarra *et al*., 2003; Koshkin *et al*., 2003; Laurents *et al*., 2003; Song *et al*., 2003; Shosheva *et al*., 2004; Sun *et al*., 2004; Toyama *et al*., 2004). After deleting some contrary and unreliable data, total 520 experimental pK_a values of six types of amino acids (Asp, Glu, Cys, Tyr, Lys and His) are used in the training set, which are listed in Table I.

Table I. The types and numbers of amino acids in training and test databases and the statistical results of Pred-pK_a calculations

Amino Acid	Asp	Glu	Cys	Tyr	Lys	His
Number	152	136	23	65	81	63
$pK_{a,\ Aver}$	3.0359	4.0151	6.0477	9.9254	10.696	6.5404
Training						
R[a]	0.9316	0.9353	0.9324	0.9596	0.8926	0.9754
RMSD[b]	0.3109	0.3974	0.4021	0.4778	0.1752	0.3042
Test[c]						
R	0.8859	0.8394	0.8653	0.8745	0.8635	0.8923
RMSD[b]	0.4357	0.4018	0.4435	0.3876	0.3439	0.4356

[a]*R*, correlation coefficient.
[b]RMSD, root-mean-squared deviation.
[c]Jackknife cross-validation test.

Table II. The symbols and implications of 17 physicochemical parameters of 20 amino acids

Parameter	Type	Physical Implication	Exponent n_l
S_L	HMLP[a]	Lipophilic surface	3
S_H	HMLP[a]	Hydrophilic surface	3
L	HMLP[a]	Lipophilic index	3
H	HMLP[a]	Hydrophilic index	3
P	Electrostatic[b]	Dipole moment	2
q	Electrostatic[b]	Net charge	1
V	Van der Waals[b]	Van der Waals volume.	12
σ	Van der Waals[b]	Diameter	6
ΔG	Desolvation[b]	Solvation free energy	3
ΔH	Desolvation[b]	Solvation enthalpy	3
$-T\Delta S$	Desolvation[b]	Solvation entropy	3
H_A	Hydrogen bond[b]	Hydrogen bond acceptor	4
H_D	Hydrogen bond[b]	Hydrogen bond donor	4
P_α	Secondary structure[b]	α-potence	3
P_β	Secondary structure[b]	β-potence	3
P_t	Secondary structure[b]	turn-potence	3
P_c	Secondary structure[b]	coil-potence	3

[a]Heuristic lipophilicity molecular potential (HMLP) (Du *et al*., 2006).
[b]The data are taken from Bava *et al*. (2004) and Toseland *et al*. (2006).

The next step is to select physical and chemical parameters of amino acids which best affect the pK_a values of residues in 3D structures of proteins. In the Pred-pK_a method, 17 physicochemical parameters of 20 amino acids are used. These 17 parameters are classified into six categories: HMLP parameters (Du *et al*., 2005a,b, 2006), van der Waals parameters, electrostatic parameters, hydrogen bond parameters and the parameters of secondary structural potency. The HMLP parameters of amino acids were developed in our previous studies (Du *et al*., 2005a,b, 2006), reflecting the lipophilic character, hydrophilic character, lipophilic surface area and hydrophilic surface area, respectively. One of the merits of the HMLP parameters is that it gives a lipophilic index and a hydrophilic index for each of the 20 amino acid side chains, describing its lipophilic moiety and hydrophilic moiety, respectively. The former reflects the hydrophobic interaction between amino acids, as well as solvation and desolvation effects; whereas the latter includes hydrogen bonding and other electrostatic interactions (Du

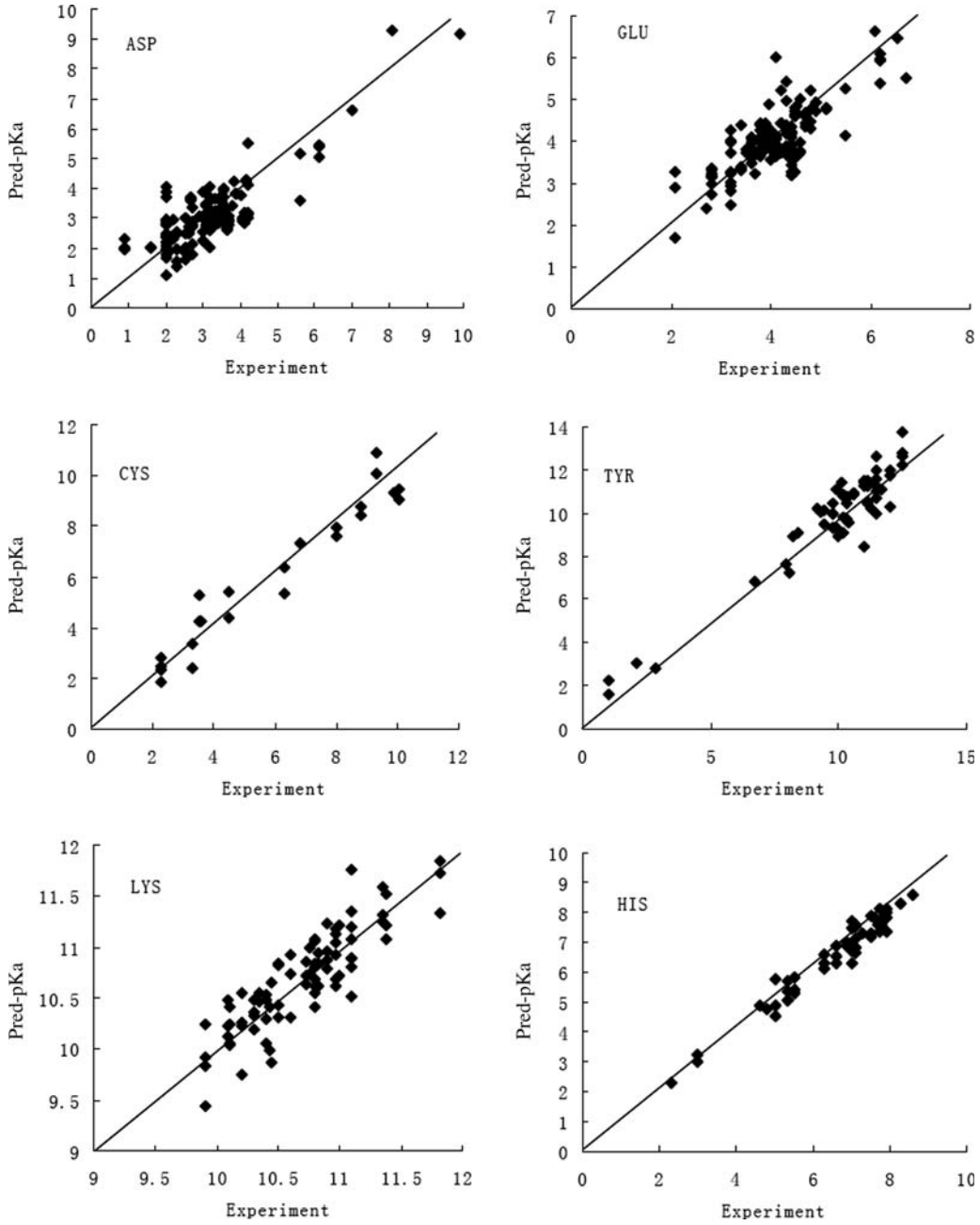

Fig. 2. The correlation relationship between the predicted pK_a in jackknife test of Pred-pK_a calculations and the experimental pK_a values of the six types of amino acids (Asp, Glu, Cys, Tyr, Lys, and His).

et al., 2005a,b, 2006). Other parameters include van der Waals volume and diameter, solvation entropy and enthalpy, dipole moment, net charge of amino acids and secondary structural potency of amino acids. The values of HMLP parameters are taken from our previous work (Du *et al*., 2006), and other parameters are taken from Gromiha *et al*. (1999, 2000). The notations and physical implications of 17 physicochemical parameters are listed in Table II.

The third step is to build the 3D data matrix $\mathbf{P}_{KML}$ for each type of amino acid according to Equations (8) and (10)–(12) using the PDB data of proteins and physicochemical parameters of amino acids. In the fourth step, the IDLS procedure is performed to determine the two sets of weight factors $\{a_\alpha^{(n)}\}$ and $b_l^{(n)}$. Then, in the fifth step, the pK_a values of query residues are predicted using Equation (18) and the values of two weight factor sets $\{a_\alpha^{(n)}\}$ and $\{b_l^{(n)}\}$.

The jackknife cross-validation test is adopted here to demonstrate the power of the current approach. In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent data set test, subsampling test and jackknife test (Chou and Zhang, 1995). In the independent data set test, although none of the proteins to be tested occur in the training data set used to train the predictor, the selection of proteins for the testing data set could be quite arbitrary unless it is sufficiently large. This kind of arbitrariness may directly affect the conclusion. For instance, a predictor yielding higher success rate than the others for a

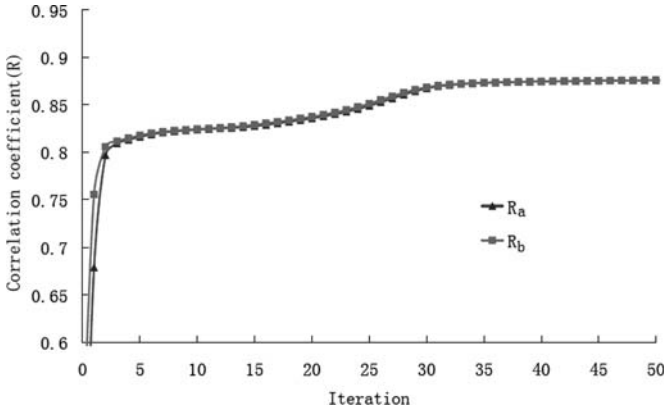

Fig. 3. The correlation coefficients between experimental and predicted pK_a values as a function of IDLS (iterative double lest square) iteration procedure. The correlation coefficient R_a is for $\{a_\alpha\}$ iteration and R_b is for $\{b_l\}$ iteration. The correlation coefficients increase with the iterations steadily and the iteration procedure converges smoothly.

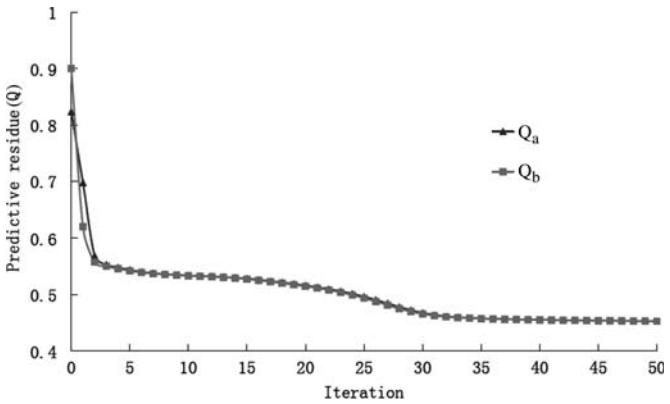

Fig. 4. The predictive residues of Pred-pK_a as the function of IDLS (iterative double lest square) iteration procedure. The predictive residue Q_a is for $\{a_\alpha\}$ iteration and Q_b is for $\{b_l\}$ iteration. The predictive residue decreases with the iterations steadily and the iteration procedure converges smoothly.

testing data set might fail to remain so when applied to another testing data set (Chou and Shen, 2008). For the subsampling test, the practical procedure often used in literatures is the 5-, 7- or 10-fold cross-validation. The problem with the subsampling examination as such is that the number of possible selections in dividing a benchmark data set is an astronomical figure even for a very simple data set [see Equation (50) of Chou and Shen (2007)]. Therefore, any practical result by the subsampling test only represents one of many possible results and hence cannot avoid the arbitrariness either. In the jackknife cross-validation, each of the statistical samples in the benchmark data set is in turn singled out as a tested protein and the predictor is trained by the remaining proteins. During the jackknifing process, both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. The jackknife cross-validation can exclude the memory effects during entire testing process and also the result thus obtained is always unique for a given benchmark data set. Therefore, of the above three examination methods, the jackknife test is deemed the most objective (Chou and Shen, 2008) and has been increasingly used by investigators to examine the accuracy of various predictors (see, e.g. Zhou *et al*., 2007, 2008; Chou and Shen, 2008; Lin, 2008; Li and Li, 2008; Zhang and Fang, 2008).

The results obtained by the jackknife tests are listed in Table I. In Table I, RMSD of predicted pK_a values of 520 residues in jackknife cross-validation test are less than ± 0.5 pH, better than other predictive methods (Davies *et al*., 2006). Figure 2 illustrates the correlation relationship between the predicted pK_a in jackknife test of Pred-pK_a calculations and the experimental pK_a values of the six types of amino acids (Asp, Glu, Cys, Tyr, Lys and His). Figure 3 shows the correlation coefficients between experimental and predicted pK_a values as a function of IDLS (iterative double lest square) iterations. The correlation coefficient R_a is for $\{a_\alpha\}$ iteration and R_b is for $\{b_l\}$ iteration. The correlation coefficients increase with the iterations steadily and the iteration procedure converges smoothly. Figure 4 shows the predictive residues of Pred-pK_a as the function of IDLS iterations. The predictive residue Q_a is for $\{a_\alpha\}$ iteration and Q_b is for $\{b_l\}$ iteration. The predictive residue decreases with the iterations steadily and the iteration procedure converges smoothly.

In order to further check the predictive power of Pred-pK_a, a comparison is carried out between Pred-pK_a and other five methods (AMBER, PARSE, MCCE, UHBD and PROPKA). For an objective comparison, a test database, containing 27 Asp residues and 12 Glu residues selected from 12 proteins, is selected, which was used by Davies *et al*. (2006). The detailed results are summarized in Table III. In Table III, the predictive pK_a values of other five methods are taken from Davies *et al*. (2006). The Pred-pK_a yields the best predictive results among the six methods.

Discussion

The Pred-pK_a model is developed on the basis of PROPKA (Li *et al*., 2005). However, it is more delicate and accurate than its original prototype. It is not a simple empirical pK_a predictive method because it is built on solid theoretical background and more physicochemical properties of amino acids are used. In the IDLS technique, the two sets of weight factors $\{a_\alpha\}$ and $\{b_l\}$ have clear physical and chemical implications. The predictive power of Pred-pK_a is affected by the selection of physical and chemical properties of amino acids. Generally speaking, more physicochemical properties are used, better predictive results are achieved. In this study, 17 physicochemical properties of amino acids are used in the Pred-pK_a calculations. However, it is possible to include more properties of amino acids in the Pred-pK_a method. In the Pred-pK_a model, the 3D structures of proteins is taken into account in the distance and orientation functions $f_l(\mathbf{r}_{k,i})$ [Equations (10)–(12)], which are designed based on the natures of physicochemical parameters and physical principles. Currently, the Pred-pK_a program ignores the possible pK_a shifts caused by bound ligands, ions and water molecules in the protein structure. This problem can be solved in the schema of Pred-pK_a program. Work including these effects is ongoing.

Table III. Comparison of Pred-pK_a with other five pK_a predictive methods

PDB[a]	Residue	AMBER	PARSE	MCCE	UHBD	PROPKA	Pred-pK_a	pK_{exp}
1A2P	Asp54	0.74	0.82	1.34	3.57	2.70	2.96	2.00
	Asp93	−0.64	−1.38	1.00	3.92	0.69	2.08	2.00
	Asp101	6.19	3.56	2.31	3.75	1.20	2.21	2.00
	Glu73	4.66	4.10	2.37	4.68	3.11	1.69	2.10
1A91	Asp7	3.99	20.29	4.17	4.04	3.87	5.16	5.60
	Asp44	6.00	20.22	5.55	4.69	4.19	3.63	5.60
	Asp61	5.52	26.20	5.01	4.33	4.01	6.65	7.00
	Glu2	4.14	7.45	4.53	4.45	4.50	4.16	5.50
	Glu37	5.15	−20.96	4.27	4.66	4.32	5.25	5.50
1BEO	Asp21	6.38	5.52	3.11	3.75	1.35	2.50	2.50
	Asp30	3.56	3.80	4.38	4.00	2.64	3.05	2.50
	Asp72	3.84	3.95	3.69	4.34	3.30	1.84	2.60
1DE3	Glu96	9.88	10.41	3.53	5.70	4.10	4.76	5.10
	Glu115	5.19	5.23	3.81	4.45	4.50	4.74	4.90
1KXI	Asp59	3.13	3.90	2.33	4.20	2.49	2.57	2.30
1LZ3	Asp18	3.79	3.94	6.77	4.05	3.19	3.35	2.70
	Asp48	3.57	3.17	5.15	3.99	2.51	1.85	2.50
	Asp66	0.23	−0.31	12.38	3.07	1.19	1.59	2.00
	Asp87	4.06	3.96	4.89	3.89	2.17	2.30	2.10
	Glu7	3.92	3.44	4.73	4.36	3.01	2.39	2.70
	Glu35	4.92	5.23	7.80	4.78	5.40	6.62	6.10
1RNZ	Asp14	9.08	7.86	3.51	4.89	−0.62	2.81	2.00
	Glu2	−1.52	−1.48	1.44	5.03	2.66	3.20	2.80
1TRS	Asp26	6.18	4.38	7.84	4.64	4.96	9.31	8.10
	Glu6	3.91	3.92	4.54	4.44	4.50	4.93	4.90
	Glu68	4.27	4.24	4.59	4.55	4.57	4.81	5.10
1TRW	Asp26	8.23	5.24	8.63	4.83	5.62	9.21	9.90
	Glu68	5.07	5.33	3.55	4.88	4.34	4.94	4.90
1XNB	Asp11	1.83	0.57	3.44	3.82	1.99	1.61	2.50
	Asp83	6.29	7.89	6.35	4.28	1.36	2.94	2.00
	Asp101	5.01	2.94	9.96	4.28	1.50	2.79	2.00
	Asp106	8.72	8.94	3.18	4.98	3.02	2.80	2.70
	Glu172	6.62	6.42	5.94	5.22	7.32	5.50	6.70
2OVO	Asp7	4.05	4.01	6.25	3.72	2.51	3.02	2.50
	Asp27	2.08	2.32	2.77	3.77	2.39	2.04	2.50
2RN2	Asp10	4.01	3.38	8.47	3.83	6.99	5.47	6.10
	Asp70	4.11	3.55	3.15	3.50	4.10	2.74	2.60
	Asp102	7.33	6.77	3.00	3.40	0.13	1.96	2.00
	Asp148	−1.10	−1.31	0.55	3.79	−0.79	2.36	2.00
RMSD	—	2.653	6.689	2.646	1.852	1.381	0.565	—

[a]The data in columns 3–7 and 9 are taken from Davies *et al*. (2006).

Conclusion

The pK_a prediction of residues in proteins is a hot topic in *Protein Engineering and Designing*. Many new approaches are developed in recent years. A comprehensive and objective comparison among these methods is very difficult because different databases and parameters are used by different authors. In this study, we only compared the Pred-pK_a with other five methods (AMBER, PARSE, MCCE, UHBD and PROPKA) (Davies *et al*., 2006). The calculation examples used in this study show that the Pred-pK_a method possesses more powerful predictive ability than other similar pK_a predictive methods. The RMSDs of predicted pK_a values using Pred-pK_a method are in the limit ± 0.6 pH, much better than other available pK_a predictive methods. However, some other methods may provide the predictive results at the same level (He *et al*., 2007; Kieseritzky and Knapp, 2008).

The foundational equation of Pred-pK_a is built on the basis of physical principle and the 3D structures of proteins. It has a strong theoretical background and better predictive power. The Pred-pK_a program gives users the freedom to select more and better physicochemical properties of amino acids. The speed, accuracy and ease-of-use of the Pred-pK_a approach, perhaps the most important conclusion of this study. Comparing with the theoretical pK_a predictive methods based on LPBE, the calculations of pK_a predictions using Pred-pK_a method only take a few minutes on common PC computer.

Funding

This work is financially supported by the National High-tech Research and Development Program (‘863’) of China under the project 2006AA020103, and by the Chinese National Science Foundation (NSFC) under the project 30970562.

References

Alexov,E.G. and Gunner,M.R. (1997) *Biophys. J.*, **72**, 2075–2093.
Alexov,E.G. and Gunner,M.R. (1999) *Biochemistry*, **38**, 8253–8270.
Antosiewicz,J., McCammon,J.A. and Gilson,M.K. (1994) *J. Mol. Biol.*, **238**, 415–436.
Antosiewicz,J., McCammon,J.A. and Gilson,M.K. (1996) *Biochemistry*, **35**, 7819–7833.
Assadi-Porter,F.M. and Fillingame,R.H. (1995) *Biochemistry*, **34**, 16186–16193.

Atkins,W.M., Wang,R.W., Bird,A.W., Newton,D.J. and Lu,A.Y. (1993) *J. Biol. Chem.*, **268**, 19188–19191.
Bartik,K., Redfield,C. and Dobson,C.M. (1994) *Biophys. J.*, **66**, 1180–1184.
Bashford,D. (1997) *An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules*. Scientific Computing in Object-Oriented Parallel Environments, **1343**, pp. 233–240.
Bashford,D. and Karplus,M. (1990) *Biochemistry*, **29**, 10219–10225.
Bava,K.A., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) *Nucleic Acids Res.*, **32**, D120–D121.
Bombarda,E., Morellet,N., Cherradi,H., Spiess,B., Bouaziz,S., Grell,E., Roques,B.P. and Mely,Y. (2001) *J. Mol. Biol.*, **310**, 659–672.
Bombarda,E., Cherradi,H., Morellet,N., Roques,B.P. and Mely,Y. (2002) *Biochemistry*, **41**, 4312–4320.
Brown,L.R., De Marco,A., Wagner,G. and Wuthrich,K. (1976) *Eur. J. Biochem.*, **62**, 103–107.
Chiang,C.M., Chang,S.L., Lin,H.J. and Wu,W.G. (1996) *Biochemistry*, **35**, 9177–9186.
Chou,K.C. and Shen,H.B. (2007) *Anal. Biochem.*, **370**, 1–16.
Chou,K.C. and Shen,H.B. (2008) *Nat. Protoc.*, **3**, 153–162.
Chou,K.C. and Zhang,C.T. (1995) *Crit. Rev. Biochem. Mol. Biol.*, **30**, 275–349.
Cohen,J.S., Griffin,J.H. and Schechter,A.N. (1973) *J. Biol. Chem.*, **248**, 4305–4310.
Concu,R., Podda,G., Uriarte,E. and Gonzalez-Diaz,H. (2009) *J. Comput. Chem.*, **30**, 1510–1520.
Consonni,R., Arosio,I., Belloni,B., Fogolari,F., Fusi,P., Shehi,E. and Zetta,L. (2003) *Biochemistry*, **42**, 1421–1429.
Czerlinski,G.H. and Dar,K. (1971) *Biochim. Biophys. Acta*, **234**, 57–61.
Davies,M.N., Toseland,C.P., Moss,D.S. and Flower,D.R. (2006) *BMC Biochem.*, **7**, 18.
Du,Q., Liu,P.J. and Mezey,P.G. (2005a) *J. Chem. Inf. Model.*, **45**, 347–353.
Du,Q., Mezey,P.G. and Chou,K.C. (2005b) *J. Comput. Chem.*, **26**, 461–470.
Du,Q.S., Li,D.P., He,W.Z. and Chou,K.C. (2006) *J. Comput. Chem.*, **27**, 685–692.
Forsyth,W.R., Gilson,M.K., Antosiewicz,J., Jaren,O.R. and Robertson,A.D. (1998) *Biochemistry*, **37**, 8643–8652.
Georgescu,R.E., Alexov,E.G. and Gunner,M.R. (2002) *Biophys. J.*, **83**, 1731–1748.
Gerratana,B., Cleland,W.W. and Frey,P.A. (2001) *Biochemistry*, **40**, 9187–9195.
González-Díaz,H. and Uriarte,E. (2005) *Bioorg. Med. Chem. Lett.*, **15**, 5088–5094.
Gonzalez-Diaz,H., Molina,R. and Uriarte,E. (2005) *FEBS Lett.*, **579**, 4297–4301.
González-Díaz,H., Vilar,S., Santana,L. and Uriarte,E. (2007) *Curr. Top. Med. Chem.*, **7**, 1025–1039.
Gonzalez-Diaz,H., Saiz-Urra,L., Molina,R., Santana,L. and Uriarte,E.A. (2007a) *J. Proteome Res.*, **6**, 904–908.
Gonzalez-Diaz,H., Saiz-Urra,L., Molina,R., Gonzalez-Diaz,Y. and Sanchez-Gonzalez,A. (2007b) *J. Comput. Chem.*, **28**, 1042–1048.
Gonzalez-Diaz,H., Prado-Prado,F. and Ubeira,F.M. (2008a) *Curr. Top. Med. Chem.*, **8**, 1676–1690.
González-Díaz,H., González-Díaz,Y., Santana,L., Ubeira,F.M. and Uriarte,E. (2008b) *Proteomics*, **8**, 750–778.
Gooley,P.R., Keniry,M.A., Dimitrov,R.A., Marsh,D.E., Keizer,D.W., Gayler,K.R. and Grant,B.R. (1998) *J. Biomol. NMR*, **12**, 523–534.
Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H. and Sarai,A. (1999) *Nucleic Acids Res.*, **27**, 286–288.
Gromiha,M.M., An,J., Kono,H., Oobatake,M., Uedaira,H., Prabakaran,P. and Sarai,A. (2000) *Nucleic Acids Res.*, **28**, 283–285.
Gustafsson,A., Etahadieh,M., Jemth,P. and Mannervik,B. (1999) *Biochemistry*, **38**, 16268–16275.
Harris,T.K. and Turner,G.J. (2002) *IUBMB Life*, **53**, 85–98.
He,Y., Xu,J. and Pan,X.M. (2007) *Proteins*, **69**, 75–82.
Hubatsch,I. and Mannervik,B. (2001) *Biochem. Biophys. Res. Commun.*, **280**, 878–882.
Ibarra,C.A., Chowdhury,P., Petrich,J.W. and Atkins,W.M. (2003) *J. Biol. Chem.*, **278**, 19257–19265.
Inagaki,F., Kawano,Y., Shimada,I., Takahashi,K. and Miyazawa,T. (1981) *J. Biochem.*, **89**, 1185–1195.
Joshi,M.D., Hedberg,A. and McIntosh,L.P. (1997) *Protein Sci.*, **6**, 2667–2670.
Kesvatera,T., Jonsson,B., Thulin,E. and Linse,S. (1996) *J Mol Biol.*, **259**, 828–839.
Kesvatera,T., Jonsson,B., Thulin,E. and Linse,S. (1999) *Proteins*, **37**, 106–115.
Khare,D., Alexander,P., Antosiewicz,J., Bryan,P., Gilson,M. and Orban,J. (1997) *Biochemistry*, **36**, 3580–3589.
Kieseritzky,G. and Knapp,E.W. (2008) *J. Chem. Comput.*, **29**, 2575–2581.
Koshkin,A., Nunn,C.M., Djordjevic,S. and Ortiz de Montellano,P.R. (2003) *J. Biol. Chem.*, **278**, 29502–29508.
Kuramitsu,S. and Hamaguchi,K. (1979) *J. Biochem.*, **85**, 443–456.
Kuramitsu,S. and Hamaguchi,K. (1980) *J. Biochem.*, **87**, 1215–1219.
Laurents,D.V., *et al*. (2003) *J. Mol. Biol.*, **325**, 1077–1092.
Li,F.M. and Li,Q.Z. (2008) *Protein Pept. Lett.*, **15**, 612–616.
Li,H., Robertson,A.D. and Jensen,J.H. (2005) *Proteins*, **61**, 704–721.
Lin,H. (2008) *J. Theor. Biol.*, **252**, 350–356.
Loewenthal,R., Sancho,J., Reinikainen,T. and Fersht,A.R. (1993) *J. Mol. Biol.*, **232**, 574–583.
Madura,J.D., *et al*. (1995) *Comp. Phys. Comm.*, **91**, 57–95.
March,K.L., Maskalick,D.G., England,R.D., Friend,S.H. and Gurd,F.R. (1982) *Biochemistry*, **21**, 5241–5251.
Matthew,J.B., Gurd,F.R., Garcia-Moreno,B., Flanagan,M.A., March,K.L. and Shire,S.J. (1985) *CRC Crit. Rev. Biochem.*, **18**, 91–197.
McNutt,M., Mullins,L.S., Raushel,F.M. and Pace,C.N. (1990) *Biochemistry*, **29**, 7572–7576.
Meadows,D.H., Roberts,G.C. and Jardetzky,O. (1969) *J. Mol. Biol.*, **45**, 491–511.
Mehler,E.L. and Guarnieri,F. (1999) *Biophys. J.*, **77**, 3–22.
Mossner,E., Iwai,H. and Glockshuber,R. (2000) *FEBS Lett.*, **477**, 21–26.
Oda,Y., Yoshida,M. and Kanaya,S. (1993) *J. Biol. Chem.*, **268**, 88–92.
Oda,Y., Yamazaki,T., Nagayama,K., Kanaya,S., Kuroda,Y. and Nakamura,H. (1994) *Biochemistry*, **33**, 5275–5284.
Oliveberg,M., Arcus,V.L. and Fersht,A.R. (1995) *Biochemistry*, **34**, 9424–9433.
Perez-Canadillas,J.M., Campos-Olivas,R., Lacadena,J., Martinez del Pozo,A., Gavilanes,J.G., Santoro,J., Rico,M. and Bruix,M. (1998) *Biochemistry*, **37**, 15865–15876.
Qin,J., Clore,G.M. and Gronenborn,A.M. (1996) *Biochemistry*, **35**, 7–13.
Reyes-Vivas,H., Hernandez-Alcantara,G., Lopez-Velazquez,G., Cabrera,N., Perez-Montfort,R., de Gomez-Puyou,M.T. and Gomez-Puyou,A. (2001) *Biochemistry*, **40**, 3134–3140.
Ruterjans,H. and Witzel,H. (1969) *Eur. J. Biochem.*, **9**, 118–127.
Schaller,W. and Robertson,A.D. (1995) *Biochemistry*, **34**, 4714–4723.
Shosheva,A., Donchev,A., Dimitrov,M., Zlatanov,I., Toromanov,G., Getov,V. and Alexov,E. (2004) *Biochim. Biophys. Acta*, **1698**, 67–75.
Snyder,G.H., Rowan,R., 3rd, Karplus,S. and Sykes,B.D. (1975) *Biochemistry*, **14**, 3765–3777.
Song,J., Laskowski,M., Jr, Qasim,M.A. and Markley,J.L. (2003) *Biochemistry*, **42**, 6380–6391.
Sorensen,M.D. and Led,J.J. (1994) *Biochemistry*, **33**, 13727–13733.
Soulere,L., Claparols,C., Perie,J. and Hoffmann,P. (2001) *Biochem. J.*, **360**, 563–567.
Sun,X., Sun,H., Ge,R., Richter,M., Woodworth,R.C., Mason,A.B. and He,Q.Y. (2004) *FEBS Lett.*, **573**, 181–185.
Szyperski,T., Antuch,W., Schick,M., Betz,A., Stone,S.R. and Wuthrich,K. (1994) *Biochemistry*, **33**, 9303–9310.
Toseland,C.P., McSparron,H. and Flower,D.R. (2006) *Nucleic Acids Res.*, **34**, 199–203.
Toyama,A., Takahashi,Y. and Takeuchi,H. (2004) *Biochemistry*, **43**, 4670–4679.
Ullmann,G.M. and Knapp,E.W. (1999) *Eur. Biophys. J.*, **28**, 533–551.
Walters,D.E. and Allerhand,A. (1980) *J. Biol. Chem.*, **255**, 6200–6204.
Yang,A.S., Gunner,M.R., Sampogna,R., Sharp,K. and Honig,B. (1993) *Proteins*, **15**, 252–265.
Zhang,G.Y. and Fang,B.S. (2008) *J. Theor. Biol.*, **253**, 310–315.
Zhang,G.Y., Li,H.C., Gao,J.Q. and Fang,B.S. (2008) *Protein Pept. Lett.*, **15**, 1132–1137.
Zhou,X.B., Chen,C., Li,Z.C. and Zou,X.Y. (2007) *J. Theor. Biol.*, **248**, 546–551.