Published in final edited form as:
Proteins. 2011 December ; 79(12): 3346–3355. doi:10.1002/prot.23098.

CALCULATION OF pK_a IN PROTEINS WITH THE MICROENVIRONMENT MODULATED-SCREENED COULOMB POTENTIAL (MM-SCP)

Jufang Shan and **Ernest L. Mehler**[*]
Department of Physiology and Biophysics, Weill Cornell Medical College of Cornell University, New York, New York 10065

Abstract

The **MM-SCP** has been applied to predict pK_a values of titratable residues in wild type and mutants of staphylococcal nuclease (SNase). The calculations were based on crystal structures made available by the Garcia-Moreno Laboratory. In the mutants, mostly deeply buried hydrophobic residues were replaced with ionizable residues, and thus their pK_a values could be measured and calculated using various methods. The data set used here consisted of a set of WT SNase for which His pK_a at several ionic strengths had been measured, a set of mutants for which measured pK_a were available and a set of 11 mutants for which the measured pK_a were not known at the time of calculation. For this latter set, blind predictions were submitted to the protein pK_a cooperative, 2009 workshop at Telluride, where the results of the blind predictions were discussed (the RMSD of the submitted set was 1.10 pH units). The calculations on the structures with known pK_a indicated that in addition to weaknesses of the method, structural issues were observed that led to larger errors (>1) in pK_a predictions. For example, different crystallography conditions or steric clashes can lead to differences in the local environment around the titratable residue, which can produce large differences in the calculated pK_a. To gain further insight into the reliability of the MM-SCP, pK_a of an extended set of 54 proteins belonging to several structural classes were carried out. Here some initial results from this study are reported to help place the SNase results in the appropriate context.

Keywords

protein pK_a values; blind pK_a prediction; microenvironment; hydrophobicity; buried fraction

INTRODUCTION

The importance of acid/base equilibria in proteins comes from the role of titratable residues (TRs) in controlling protein structure, solubility and function, especially in cases where the pK_a values show large shifts from their solvent values ($\Delta pK_a > 1$). Such residues are frequently associated with protein stability [1-3], or where the pK_a value shifts into the biological pH range, the residue is often involved in the protein's function[4-11]. Because of this a strong demand has developed for methods that can predict pK_a values with accuracy and provide physical insight into the macromolecular forces that govern the protein's properties. Protein structure is characterized by a complex mosaic of variable dielectric regions [12] that has provided nature with a tool to evolve specialized protein architectures able to induce large changes in the properties of embedded amino acid residues. However, it

[*] Corresponding Author: elm2020@med.cornell.edu, Phone: 212-746-6365.

is just this complexity of protein structure that has made it difficult to develop methods that are uniformly reliable in predicting pK_a values of the ionizable residues in all proteins, as is documented in two recent reviews [13,14] where a number of methods are compared.

The approach we have adopted combines the electrostatic theory of Lorentz, Debye, Sack and Onsager (LDSO) [15,16] with a variational approach to optimize the total electrostatic free energy with respect to the distribution of the ionization charge[12,17]. To help account for the structural complexity alluded to above, we have developed an approach that allows the electrostatic properties of the TRs to be modulated by the hydrophobic properties of the local environment embedding the ionizable group. The method is named the Microenvironment Modulated–Screened Coulomb Potential (MM-SCP) approximation. After a brief review of the methodology we first present results from calculations on wild type (WT) and mutants of SNase. First a series of studies were carried out on WT and mutants for which the pK_a values were already published [4,18-22] (see data from the Garcia-Moreno laboratory) to obtain insight into how the SNase system behaves, and subsequently blind predictions were submitted to the pKcoop prior to the summer, 2009, meeting at Telluride. However, the overall reliability of a given method cannot be determined from the results of a single protein, but requires an extended sample of proteins representing different structural types. To test the MM-SCP, and hopefully to use the results to develop an improved, more reliable algorithm, we have constructed a data set of 334 ionizable residues of measured pK_a values from 54 proteins, and in the second part of this report we present some initial results from our studies of this data set.

METHODS

Calculation of pK_a in Proteins with the MM-SCP

Most methods for calculating acid-base equilibria in proteins are based on a thermodynamic cycle[23] of the form

$$\begin{array}{ccccccc} & & & 2.303RT\ \mathrm{pK_a(s)} & & & \\ & & \mathrm{AH(s)} & \Longleftrightarrow & \mathrm{A^-(s)+H^+(s)} & & \\ \Delta\mu_{\mathrm{s,p}}(\mathrm{AH}) & & \Updownarrow & & \Updownarrow & & \Delta\mu_{\mathrm{s,p}}(\mathrm{A^-}) \\ & & \mathrm{AH(p)} & \Longleftrightarrow & \mathrm{A^-(p)+H^+(s)} & & \\ & & & 2.303RT\ \mathrm{pK_a(p)} & & & \end{array} \quad (1)$$

where p and s refer to the protein and the solvent, respectively. Thus the pK_a of a TR in the protein can be calculated from a reference $pK_a(s)$ and the additional changes in free energy that arise when the group is transferred from the solvent into the protein. Here only electrostatic contributions are included so that the pK_a of the group in the protein is calculated from

$$\mathrm{pK}_a(\mathrm{p})=\mathrm{pK}_a(\mathrm{s})+(w_A^{\mathrm{int}}+\alpha_A\Delta w_A^{tr})/2.303\ RT \quad (2)$$

where α_A is a linear scaling factor to be discussed below, w_A^{int} is the interaction free electrostatic energy of the charged group in the field of *all* the other groups in the protein, and Δw_A^{tr} is the change in self-energy on transferring the group from the solvent into the protein (for details see Reference [12]). The interaction energy, w_A^{int} and transfer energy (desolvation/resolvation), Δw_A^{tr}, are defined in such a way that the contribution of the M ionizable groups to the total electrostatic free energy w of the system can be expressed as

$$w=\sum_{A=1}^{M}(w_{A}^{\mathrm{int}}+\alpha_{A}\Delta w_{A}^{tr}) \tag{3}$$

In this method the interaction energies between non-titratable groups are constant, and can be set to zero. In most cases the interaction energy contributes to stabilizing the system, whereas the transfer energy is positive in almost all cases because the charge is transferred from a hydrophilic environment (the aqueous solvent) to a more hydrophobic environment (the protein).

In most pK_a calculation methods the equilibrium charge state is calculated from the distribution of the charge over the 2^M ionization microstates using a Monte Carlo approach. The method developed in references [12,17] proceeds in a different way: the assignment of the ionization charge over the atoms of the protonatable moiety is determined through a variational optimization of the total electrostatic free energy (i.e., $\delta w = 0$), thus allowing the TR to respond to the protein-solvent environment. The total variational ionization charge of each group is coupled to the pH via the Henderson-Hasselbach equation.

To derive the variational equations for assigning the titration charge, the partial charge of an atom *a* in a group *A* is defined as

$$q_a=(1-\theta_{A})q_a^n+f_a q_a^o \tag{4}$$

where, q_a^o are fixed initial partial charges ($\neq 0$ for titratable groups), f_a are scaling factors that will be optimized to determine a stationary point of *w, and* θ_A *is the fraction of A in the charged state*. For simplicity, the group subscript *A* will be omitted, thus $q_A = \Sigma_a q_a$, where *a* runs over all the atoms in *A*; q_a^n is the partial charge from the neutral group, i.e., $\sum q_a^n=0$, and θ_A, is zero for neutral groups. Thus the first term on the r.h.s. of Eq.(4) is the contribution to the charge q_a from the neutral form of the protonatable group, while the second term represents the contribution from the ionization charge.

Analysis of local environments around TR known to have large shifted pK_a favoring the neutral state shows that many of the nearest neighbor residues around these TRs are hydrophobic, and the greater the shift in pK_a value the more hydrophobic these residues tend to be [24]. To allow an accounting of this hydrophobicity in the calculation of pK_a a quantitative representation of the degree of hydrophobicity/hydrophilicity (Hpy) of the local region surrounding the ionizable group is needed. The approach taken makes use of the “Rekker Fragmental Hydrophobic Constants” [25] and has been described elsewhere [12]. Rekker’s approach assigns hydrophobicity parameters to small chemical functionalities (in some cases individual atoms), and Hpy values are calculated by summing the Rekker fragmental constants for all fragments in the nearest neighbor shell (the microenvironment) around the TR in the protein. Rekker fragmental constants are completely independent of any protein properties, although they can be used to derive a hydrophobicity scale for the amino acid residues [25-27].

The total hydrophobicity/hydrophilicity ($THpy_A$) of the microenvironment of a group *A* is calculated from two contributions and the solvent exposed fraction, ξ_A,

$$\mathrm{THpy}_{A}=(1-\xi_{A})\ \mathrm{Hpy}_{A}+\xi_{A}\ \mathrm{Hpy}_{A}^{o} \tag{5}$$

where Hpy_A is the contribution from the protein and $Hpy_A{}^o$ is the contribution from the solvent. Finally, since THpy is an extensive quantity, it is normalized in the form

$rHpy_A = THpy_A / Hpy^o_A$, where rHpy values range from about 1 (the value of pure water) to about -0.4. This shows that increasing hydrophobicity correlates with decreasing rHpy, with negative values indicating extremely hydrophobic microenvironments. Once all the $rHpy_A$ have been evaluated, they are used to assign the optimal screening function and the scaling factor α_A. In most cases α_A is one, but for hydrophobic microenvironments this is insufficient to yield the correct energy penalty for transferring a charged group from water into the protein. These microenvironments are very different from the typical hydrophilic-to-slightly-hydrophobic microenvironments that surround most titratable groups, and tend to have little effect on their pK_a values. These latter microenvironments, however, were used to determine the values of the Born radii required to calculate the self energies and therefore these radii cannot account for the unusual local environments that tend to induce large pK_a shifts. Calibration of the scaling factor was based on highly-shifted experimental pK_a values of several proteins [24]. The results indicated that $\alpha_A = 1$ for rHpy > 0.1; $\alpha_A = 2$ for $0.02 < rHpy < 0.1$; and $\alpha_A = 3$ for $rHpy < 0.02$. It is noted that α is treated as a step function which may be problematic under certain conditions (see below).

Program input parameters

The MM-SCP program is freely available. Please contact ELM (mailto:elm2020@med.cornell.edu). To allow the user maximum control over the parameters that control the calculation, the code has been written so that the user can change these parameters to better describe the physics of the system being studied. Although this approach may make using the program more difficult it does allow for the user’s insight to be reflected in the outcomes of the calculations. In any event we have supplied default values for all the input parameters. To illustrate the types of input parameters that can be modified a few are discussed here:

1. IONS – Sets the ionic strength of the system; the default value is 0.15 M.
2. FDAM – Controls the rate of updating the charge distribution scaling factors. Thus, for the n’th iteration $f^n = \lambda f(new) + (1-\lambda)f^{n-1}$ where λ is the damping factor. The default value of λ is 0.35, although smaller values seem to lead to faster convergence in most cases.
3. NBCH – van der Waals distance damping. NBCH defines the distance threshold for damping interaction energies between charges that are too close. The idea here was that for charges closer together than their van der Waals distance, the point charge model breaks down. Therefore the interaction energy between such charges is damped by a scaling factor. Most results obtained with the MM-SCP suggest that NBCH should have a value around 3 Å with a scale factor around 0.1. Thus, the calculations reported here use these values for NBCH and the scaling factor unless noted otherwise.

Preparation of data

PDBs were prepared by removing co-factors, ligands, ions, waters and detergents, and then applying HBUILD (PAR 22 forcefield [28]) to add all hydrogens. No subsequent minimization of the crystal structure coordinates was carried out. Partial charges from the PAR 19 [29] data set were used for all residues except the neutral forms of the ionizable residues for which charges from PAR22 [28] were used. The default reference pK_a values for Asp, Glu, His, Tyr, Lys, Arg, N- and C-terminal residues used in all calculations were 4, 4.4, 6.3, 10, 10.4, 12.0, 7.5 and 3.8, respectively. It is noted that the parameters defined in reference[12] were used here.

The extended data set (Supplemental Table I lists all proteins used in the data set and their RMSD) required some additional considerations to obtain a set of experimental pK_a values of reasonably uniform quality. The following pK_a values were excluded from the calculations: residues with several different experimental pK_a values, pK_a measured during acid-denaturation, residues predicted with large errors by multiple methods [13,30] and residues with poor X-ray structure (all pK_a values are given in Supplemental Table II). The final dataset has 334 pK_a values from 54 proteins belonging to five different structural classes according to SCOP (Structural Classification of Proteins) [31] (see supplemental Table I).

RESULTS AND DISCUSSION

pK_a calculations on WT and mutant SNase

Histidines—In WT SNase (pdb code 1stn) there are four His, and their pK_a values have been measured at several ionic strengths. In earlier calculations the pK_a of the three His (positions 8, 46, and 121) in PHS SNase had been calculated [24] and the values were all within around 0.5 pK units of the experimental values showing that scaling of local hydrophobic environments (discussed above) had no effect on other TRs. Those results suggested that WT SNase be tested in the same way at all four His positions in the sequence and at various ionic strengths (I) (0.01, 0.02, 0.1,0.5, and 1.5 M). First calculations were carried out at I = 0.1 M with NBCH = 1.9 Å and the damping factor (β) was 0.35. The pK_a value of H46 was 2.47 (error of -3.40) while the values for the other His were within 0.7 units of the experimental values. Repeating the calculation with NBCH = 3 Å and β = 0.1 resulted in the pK_a value of H46 to increase to 3.64 (error = -2.22) whereas the pK_a values of the other His were negligibly affected. Such sensitivity to the value of NBCH suggests that the structure around H46 involves steric clashes. To alleviate these, a local minimization (LM) was carried out on residues 45, 46 and 47. The subsequent pK_a calculation yielded a value of 5.75 (error = -0.11) with no changes in the values of the other pK_a. Finally, as a control, LM was carried out on all histidines (at I = 0.1 M), resulting in all errors <1. Calculations were then carried out using the four different ionic strength values (data not shown). A cutoff value of 3 Å for NBCH with a damping of 0.1 seem to be reasonably optimal values. The calculated pK_a values had an overall RMSD to the experimental values of 0.34 and the errors ranged from -0.70 to + 0.65. Thus the four His pK_a were quite well reproduced at all ionic strengths. Nevertheless, as shall be seen in the second part of this report, it cannot be concluded that these good results imply the same level of accuracy will be achieved for calculation of His pK_a values in other proteins.

Mutants with known pK_a values—Table I lists calculated pK_a values for a number of SNase mutants where experimental pK_a were known at the time of calculation. The mutants are listed in the Garcia-Moreno lab database (Table 2) and are included here to illustrate some issues that may arise when using coordinates obtained from crystal structures. Entry *1* in the Table I is the V66D mutant of the ΔPHS construct. It is seen that the residue is deeply buried in the protein and is embedded in a hydrophobic local environment. Nevertheless the calculated pK_a value is in error by 1.3 pH units. The LM procedure described above was applied and the error in the pK_a value now was reduced to 0.1 (*2* in the Table I). In both cases the microenvironments are hydrophobic, but more so in the LM structure (Table I). Nevertheless the transfer energy scaling factor is the same for both structures. Thus, it cannot be concluded that the issue here is necessarily structural, but may be due to instabilities in the iterative procedure which only appears under certain structural conditions.

Entries *3* and *4* in Table I report the results for the I92E mutant where the crystallography of the former was carried out at cryogenic temperature (CRT) while the latter was carried out at room temperature (RT) [32]. It is seen that the structural properties of the two crystal structures are very similar and that the microenvironments are extremely hydrophobic. The pK_a values were shifted upwards by around 4 pK_a units in good agreement with experiment. Entry *5* in Table I shows that the calculated value of the pK_a of I92K is shifted downwards by about 5 pK units, as expected from its highly hydrophobic microenvironment, and in agreement with experiment.

Entries *6*–*9* report results for the K and E mutants of L38. The results show that the influence on structure of mutations with different charge can be quite significant. For both mutants the TRs are essentially 100% buried, but for the K mutant the microenvironment is nevertheless relatively hydrophilic resulting in a value of one for α. This results in a net energy (see Eq.(2)) of around 0.6 Kcal/mol depressing the pK_a values by about 0.4 pH units. LM depresses the pK_a value somewhat more, but they are only slightly shifted in both cases from their solution values, in reasonable agreement with the measured threshold value. For the E mutant the situation is somewhat different: It is seen that the pK_a value of the LM structure is in good agreement with experiment, but without the LM treatment the error is -1.5. The BF values of the two structures are identical, but the rHpy value of the LM structure is 0.07 as compared to 0.14 for the nonLM structure. Thus LM has altered the microenvironment in such a way that its hydrophobicity is sufficiently increased to assign a value of 2 to α, which is reflected in an increase in the pK_a shift from 0.5 for the nonLM case to 1.6 for the LM case. It is important to note that in this particular case of LM, it is not just the structural optimization that is responsible for the improvement, but the increase in hydrophobicity due to small structural changes in the residues defining the microenvironment, which reassigns the value of α.

The coordinate file for I72E contains two coordinate sets for E72, which are sufficiently different to effect sizable changes in the local environments for E72. With the A coordinates E72 is embedded in a relatively hydrophilic microenvironment and α was assigned the standard value of 1 while the B coordinate set defines a strongly hydrophobic local environment with α = 3. It is seen that the pK_a value of the A coordinates is shifted upwards but not enough, while the B coordinates shift the pK_a upwards too much. The origin of the relatively large shift in rHpy is the difference in solvent exposed surface area. Although this difference is fairly small (0.142 and 0.081, from the A and B coordinates, respectively) the effect on rHpy is large because the hydrophilicity of water is so large. Therefore a relatively small change in solvent exposed surface area has a concomitantly large effect on the value of THpy, and rHpy (see Eq.(5), and the immediately following discussion). It would be of interest to carry out MD simulations on these two systems to determine if both structures lead to the same final pKa value.

In blind predictions nothing is known about the pK_a value and it is tempting to take the average of the two values for I72E which in this case yields 7.3. LM does not change the results and yields an average value of 6.9. For the results given in Table I the LM structure gives an improved value of pK_a for all but one case, namely L38K where the pK_a calculated for the LM structure has a somewhat larger error than from the nonLM structure. This suggests that until a reliable method is available to differentiate between the LM and nonLM structures, using the average value might be a reasonable compromise. This yields a value of 7.3 for V66D, 9.7 for L38K, and 6.1 for L38E and the error is <1 for all these cases.

Blind predictions—Table II lists the 11 blind predictions that were submitted to the pKcoop for the 2009 workshop (the pK_a calculations discussed in the above sections were carried out on systems where the pK_a values were known). For each mutant the pK_a was calculated with nonLM and LM coordinates. For eight mutants the difference between the two coordinate sets were negligible (<0.2) and the nonLM pK_a values are given in Table II, but for three cases (T62K, L103K and V104K) the differences were larger with the pK_a improved for T62K, but worsened for the last two mutants (a result that, of course, was not known at the time of submission). Seven (of eleven) mutants were in error by <1.0 pK units, while one had an error > 2 pK units. A109R and A90R: In spite of the fairly high BF the rHpy values indicate a moderate hydrophobic microenvironment that shifts the pK_a values slightly downward in both mutants resulting in pK_a values in agreement with experiment. I72K: The calculated pK_a of this mutant is in error by 2.6 pK units which is the largest for the blind data set. It is noted that both BF and rHpy indicate a substantially solvent exposed hydrophilic microenvironment which is also seen in the crystal structure. Thus the downward shift of the measured pK_a may indicate that around 72K the crystal structure disagrees with the solvent structure. In any event given the solvent exposure and resulting rHpy, the calculated pK_a is typical for the MM-SCP, and the source of the error is not clear. L25K and L25E: Both mutant TRs are deeply buried in the protein with rHpy values indicating strongly hydrophobic microenvironments, and both residues show strong shifts in their pK_a values as expected. The LM structure of L25E yields a pK_a of 8.5, slightly closer to the experimental value than the value from the nonLM coordinates. L36K: The mutant residue is deeply buried in the protein with a strongly hydrophobic local environment. The calculated downward shift of the pK_a is in reasonable agreement with the observed value. T62K: For this mutant the pK_a value of the LM coordinate set is closer to the experimental value than the nonLM result. The pK_a values for the nonLM and LM coordinate sets are 6.5 and 8.1, respectively, and it is seen that the latter value is close to the experimental value. Because it was not possible to know this, the average value of the pK_a was submitted. Further analysis of the results suggests again that the algorithm that assigns the value of the scale factor needs to be refined. Y91E: In this mutant, the Glu is well buried in the protein in a hydrophobic microenvironment. The calculated pK_a is in good agreement with experiment. L103K: Although the Lys is deeply buried in the protein, the rHpy value indicates only a weakly hydrophobic local environment. The pK_a values of the nonLM and LM structures are 9.8 and 10.5, respectively, so that the latter value is worse than the former. However, analysis of the titration curves indicated that the nonLM titration was more stable than the LM titration. Therefore we submitted the nonLM pK_a value. V104K: The nonLM and LM pK_a values are 7.9 and 9.4 units, respectively, and Lys is completely buried in the protein, but the values of rHpy again suggest that α is not correctly assigned. We submitted the average value of the two pK_a for the blind prediction. L125K: This mutation is buried in the protein in a fairly strong hydrophobic microenvironment. The calculated pK_a value is in good agreement with the measured value.

The results of these blind predictions suggest that the MM-SCP in its present form is reasonably reliable, but definitely needs further improvement. What the results support is that large shifts in pK_a due to the ionizable residue being embedded in hydrophobic microenvironments can be reasonably well predicted. Whether large shifts due to other structural adaptations in a protein can also be predicted remains to be seen. LM is an attractive approach to relieve steric clashes in a TR due to its simplicity. Clearly, it needs further refinement to become a viable method for preprocessing structures for pK_a prediction. One likely source of error in the present method of minimization is the absence of solvent except via a linear distant dependent dielectric screening. Recently, we have started analyzing the MM-SCP using a large data set of proteins with many measured pK_a values. To start elucidating some of the questions mentioned above we report some initial results from this expanded analysis.

pK_a calculations on an extended data set of proteins

Overall the protein RMSDs of pK_a prediction errors (Supplemental Table I) range from a low of 0.22 to two values > 3 (although one of them decreased to < 1 after LM); 35 (of 54) proteins have an RMSD < 1.0 whereas the RMSD of six proteins is > 2. This clearly shows that it is not possible to evaluate the reliability of a particular method for calculating pK_a on the basis of one, or even a few proteins. At the same time it should be recognized that as the methods improve smaller data sets should be sufficient to test the method's reliability. The RMSD for the entire set of 334 TR in 54 proteins is 0.96, and as summarized in Table III, around 18% of the calculated pK_a are in error by > 1 pH unit. In Table IV the proteins have been collected by their SCOP classes [31] and the RMSDs have been calculated for each class (Table IV). The results show some differences in class RMSD values. It is seen that overall small proteins (class *g*) have the smallest errors in their calculated pK_a values. This is not surprising since for these proteins most TR's will be on or near the protein surface and the crystal structures are apt to be at higher resolutions. However, for classes *a* – *d* there are substantial differences in the RMSD values. Thus class *d* (a+b) has the lowest RMSD whereas *c* (a/b) has the largest value; classes *a* (all α helix) and *b* (all β sheet) have similar intermediate values. Although the sample sizes of the classes (32 residues for *g* to a maximum of 100 residues for *d*) may not be large enough to be statistically significant, they are nevertheless large enough to pose intriguing questions regarding the origin of these differences in class RMSD. Thus, can these differences be used directly to improve the reliability of pK_a calculations, or do they indicate an underlying higher level of complexity? Since the main purpose of including this section in the manuscript is to analyze the sources of errors the following sections discuss various aspects of this issue.

Error vs. Shifts—It has long been apparent that pK_a values that are strongly shifted from their reference values are also the pK_a most prone to large errors. This trend is also seen in the MM-SCP algorithm as shown in Table III. Thus for the 237 residues where the measured pK_a are shifted by less than one pH unit only 14 residues are in error by more than one unit. In contrast, for the 97 residues that are shifted by more than one unit about half of the calculated values are in error by more than one pH unit. The SCOP class RMSDs show the same behavior: thus for shifts < 1 pH unit the class RMSDs are all < 1, but for larger shifts all the class RMSD increase with the class *c* value being more than twice as large as the next largest RMSD (class *b*). The relationship between shifts and errors, is plotted in Figure 1, which shows that large positive shifts are associated with negative errors and vice versa. Note also that there are one Glu and one His with zero shifts that nevertheless are substantially in error which could be resolved by LM and there are a number of cases with shifts between 1 and 2 pH units where the calculated pK_a from the LM structure are < 1 from the measured value.

Error vs. Residue Type—Table V shows that the residue type with the biggest RMSD is Tyr, because there is a large fraction of Tyr residues with big shifts in our dataset (15 out of 24) while only about half of them were predicted with small errors (Figure 1). Since there are only 24 Tyr in the data set the results for this residue are of least statistical significance. The second biggest RMSD is for His residues with 34% having large errors. In detail, 66% of His with large shifts have large errors and even for His with small shifts, 16% were predicted with large errors. Moreover, His has the highest RMSD for small shifts (Table V). Thus, in the MM-SCP approach the main problem of pK_a prediction lies in addressing the errors in the calculated values of His. As the reference pK_a for His (6.3) is close to physiological pH, improved accuracy of His pK_a predictions would have a large impact on the overall accuracy of the method, and on understanding/probing the biological questions involving His residues. Thus the extended data set suggests much greater difficulty in calculating pK_a of His than was implied from the WT SNase calculations.

Error vs. Environment (BF & rHpy)—It is also of interest to ask if there is a relationship between error size and degree of embedding of the titratable moiety in the protein, which is shown in Figure 2. It is noted that most residues are located in the region with BF < 0.7, i.e., near the surface of the protein, and most such residues are in the acceptable error range or are in error by < 2 pH units. There are less residues in the region 0.7 < BF < 1.0 (83 out of 334), but it is also seen that the density of points with acceptable calculated pK_a is much lower in this latter region than in the smaller BF region, and many of the residues with large errors are also found in the large BF region. Comparison of these regions suggests that pK_a with large shifts are often found in the high BF region. This is probably to be expected since the evolution of protein architecture that leads to residues with aberrant properties requires access to the entire residue, not just a small part with the rest immersed in solvent. The effect of the hydrophilicity/hydrophobicity of the environment on pK_a prediction with the MM-SCP was analyzed by plotting errors vs. rHpy (Figure 3). Only 8 residues of 334 are embedded in the formally defined hydrophobic environment (see Methods) and 6 were predicted with error <1 unit. There seems to be no obvious relationship between the error and rHpy. At the same time Figure 3 shows that some large errors concentrate around rHpy between 0.1 and 0.25 suggesting again that tweaking of the algorithm for assigning the hydrophobicity scaling factor might improve the MM-SCP algorithm, as shown in our earlier study [24]. Interestingly there is no His in a hydrophobic environment indicating the error in calculating pK_a for His might arise from improper treatment of interaction energies (discussed further below).

Other Sources of Errors in the MM-SCP Approximation—As discussed above, MM-SCP treats the hydrophobic environment relatively well thus another major source of error might be insufficient treatment of the interaction energies. Structural analysis shows that some residues with large errors are involved in specific interactions, e.g., some His form ionic H-bonds and cation-π interactions. Several studies have suggested the importance of proper positioning of protons and optimizing H-bond network in pK_a predictions [33-36]. It was found that most of the residues predicted with large errors have some H-bonds, but a lot of residues with small errors also H-bond to other residues. Further analysis indicated that TR with large errors were enriched with certain types of H-bonds, and found that 8 of 27 such His form ionic H-bonds with Asp, Glu or the C-terminal (identified by using a PD – PA distance cutoff of 3.2 Å), which is much higher than that for His predicted with small errors (2 of 53), suggesting the possible insufficient treatment of such H-bonds in the MM-SCP algorithm. Interestingly, His predicted with large errors also form face-face and face-edge interactions with Phe or Tyr more often than those predicted with small errors (6 of 27 vs 1 of 53). Such interactions were identified if either N in the indole ring of His is within 4 Å of at least three ring atoms in Phe or Tyr. In all these cases, the pK_a is underestimated, which may be due to the interaction energy of the ionized form of His with Phe or Tyr, i.e., cation–π interactions, not being properly accounted for by the MM-SCP. Strong ionic H-bonds can also affect the pK_a of Asp as is demonstrated by several cases where improper treatment of strong ionic H-bonds may be responsible for large errors. For Tyr, all six residues with positive big errors have pK_a downshifted by at least 1.7 units, which is not reproduced in the calculated pK_a values.

Improving Accuracy by Relieving Steric Clashes—As mentioned above, LM resolves some large errors by removing steric clashes. Thus we searched for systematic ways to identify such clashes. First we searched for pK_a values predicted differently with NBCH=1.9 and NBCH=3. Visual inspection of these residues that have different calculated pK_a (>0.5 unit) shows that most of them (34 of 43) are involved either in steric clashes (10) or non optimal H-bonding (24), which is consistent with the observation that crystal packing induced structural artifacts can lead to large errors [34]. Notably, LM described in the previous section (SNase) improves the pK_a prediction for 12 residues of 43 selected by the NBCH test by at least 0.5 unit. For example, the error for His119 in 1YMB decreased from -3.97 to

-0.62 in magnitude, and the error for Glu78 in 1HV1 decreased from 3.49 to 0.87. However, NBCH does not seem to be an effective way to identify steric clashes because visual inspection is required to differentiate steric clashes from strong H-bonds with short PA-PD distances. Thus we tried to use different criteria to identify steric clashes and found that a distance cutoff of 1.5 Å not only identified all the steric clashes suggested by NBCH but also one residue not picked up by the NBCH criteria (His33 in 1HRC).

CONCLUSIONS

We have applied the MM-SCP algorithm to predict pK_a for His in WT SNase at several ionic strengths, and mutant SNase where the pK_a value was known and a set of 11 mutants that were predicted blind. To gain further insight into the reliability of the MM-SCP approach, it was applied to calculate pK_a for an extended data set of 334 TR in 54 proteins. The RMSD of the blind prediction data set was 1.1 pH units, which can be compared to the value of 0.96 obtained for the entire extended data set. In both cases almost all values are in error < 2 pH units, but there remain a few cases with errors > 2. These results seem to well represent the status of the MM-SCP at its present state of development. The RMSD for all the entries in Table I can be calculated by taking averages of the LM and nonLM values leaving six unique pK_a values. The RMSD for these is 0.54, which, compared with the above results, is a strong reminder that caution must be exercised when evaluating a method on the basis of a small data set and *a priory* known experimental values. Nevertheless, the RMSD of 0.96 decreased to 0.88 after LM of thirteen residues identified either by the criteria of NBCH or the distance cutoff. Analysis of the extended data set showed that most errors come from residues with large pK_a shifts, and that His and Tyr appear to be most difficult to predict correctly with the MM-SCP algorithm. Analysis of the energy components showed that the current MM-SCP algorithm accounts for the transfer energy (related to the local environment) relatively well, but the treatment of interaction energies still needs improvement. In particular, the treatment of ionic H-bonding and cation-π interactions needs further development. In addition the evaluation of the scaling factor, α, is being modified to make the assigned values less sensitive to small changes in the value of rHpy.

In spite of the above issues the MM-SCP algorithm appears to be accurate enough to allow the identification of errors that may be due to factors unrelated to the reliability of the method. The calculations reported in this work were all carried out on crystal structures and in several cases the calculated pK_a were in error by ~1-2 pH units, some of which were resolved by carrying out LM (see Table I). In other cases both the nonLM and LM calculations were in error, but in opposite directions, which suggested to take the average of the two calculated values, which seemed to work fairly well. In any event errors that can be resolved by LM of the TR may have a number of causes two of which are artifacts in the crystal structure or incorrectly placed hydrogen atoms, or a combination of both. It is, of course, clear that under blind prediction conditions there would be no way to differentiate between all these possibilities so that for blind predictions using the average pK_a value seems sensible. It would be of interest if molecular dynamics (MD) could resolve such issues, but it needs to be kept in mind that using MD would spoil the computational efficiency of the method, and it would probably be more efficient and accurate to use an MD based approach directly.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. Bertrand Garcia-Moreno and his laboratory for making available the crystallographic data on which this work is based. Computational support was provided by the National Science Foundation Terascale Computing System at the Texas Advanced Computing Center (TG-MCB080118N). The authors also acknowledge access to the computer facilities at the Institute of Computational Biomedicine (ICB) of Weill Cornell Medical College of Cornell University. The support of NIH grant number R01 DA015170 is gratefully acknowledged.

References

1. Pace CN, Grimsley GR, Scholtz JM. Protein Ionizable Groups: pK Values and Their Contribution to Protein Stability and Solubility. Journal of Biological Chemistry. 2009; 284(20):13285–13289. [PubMed: 19164280]
2. Whitten ST, Garcia-Moreno EB. pH Dependence of Stability of Staphylococcal Nuclease: Evidence of Substantial Electrostatic Interactions in the Denatured State. Biochemistry. 2000; 39(46):14292–14304. [PubMed: 11087378]
3. Yan ECY, Kazmi MA, De S, Chang BSW, Seibert C, Marin EP, Mathies RA, Sakmar TP. Function of Extracellular Loop 2 in Rhodopsin: Glutamic Acid 181 Modulates Stability and Absorption Wavelength of Metarhodopsin II. Biochemistry. 2002; 41(11):3620–3627. [PubMed: 11888278]
4. Castaneda CA, Fitch CA, Majumdar A, Khangulov V, Schlessman JL, Garcia-Moreno BE. Molecular determinants of the pK_a values of Asp and Glu residues in staphylococcal nuclease. Proteins. 2009; 77(3):570–588. [PubMed: 19533744]
5. Bartik K, Redfield C, Dobson CM. Measurement of the Individual pK_a Values of Acidic Residues of Hen and Turkey Lysozymes by Two-Dimensional NMR. Biophys J. 1994; 66(4):1180–1184. [PubMed: 8038389]
6. Kuramitsu S, Hamaguchi K. Analysis of the Acid-Base Titration Curve of Hen Lysozyme. J Biochem. 1980; 87(4):1215–1219. [PubMed: 6771251]
7. 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(17):5275–5284. [PubMed: 7909691]
8. Oda Y, Yoshida M, Kanaya S. Role of Histidine 124 in the Catalytic Function of Ribonuclease HI from *Escherichia coli*. J Biol Chem. 1993; 268(1):88–92. [PubMed: 8380173]
9. Joshi MD, Hedberg A, McIntosh LP. Complete measurement of the pK_a values of the carboxyl and imidazole groups in *Bacillus circulans* xylanase. Protein Science. 1997; 6(12):2667–2670. [PubMed: 9416621]
10. 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(7):1655–1674. [PubMed: 10677214]
11. Shaw RW, Hartzell CR. Hydrogen ion titration of horse heart ferricytochrome c. Biochemistry. 1976; 15(9):1909–1914. [PubMed: 5119]
12. Mehler EL, Guarnieri F. A Self-Consistent, Microenvironment Modulated Screened Coulomb Potential Approximation to Calculate pH Dependent Electrostatic Effects in Proteins. Biophysics J. 1999; 77(1):3–22.
13. Stanton CL, Houk KN. Benchmarking pK_a Prediction Methods for Residues in Proteins. J Chem Theory and Comp. 2008; 4(6):951–966.
14. Lee AC, Crippen GM. Predicting pK_a. Journal of Chemical Information and Modeling. 2009; 49(9):2013–2033. [PubMed: 19702243]
15. Hassan SA, Mehler EL. From Quantum Chemistry and the Classical Theory of Polar Liquids to Continuum Approximations in Molecular Mechanics Calculations. Int J Quant Chem. 2005; 102(5):986.
16. Mehler, EL. The Lorentz-Debye-Sack Theory and Dielectric Screening of Electrostatic Effects in Proteins and Nucleic Acids. In: Murray, JS.; Sen, K., editors. Molecular Electrostatic Potential: Concepts and Applications. Vol. 3. Amsterdam: Elsevier Science; 1996. p. 371-405.

17. Mehler EL. Self-Consistent, Free Energy Based Approximation to Calculate pH Dependent Electrostatic Effects in Proteins. J Phys Chem. 1996; 100(39):16006–16018.

18. Garcia-Moreno B, Chen LX, March KL, Gurd RS, Gurd FRN. Electrostatic Interactions in Sperm Whale Myoglobin. J Biol Chem. 1985; 260(26):14070–14082. [PubMed: 4055771]

19. Garcia-Moreno 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(1-3):211–224. [PubMed: 9127946]

20. Harms MJ, Castaneda CA, Schlessman JL, Sue GR, Isom DG, Cannon BR, Garcia-Moreno B. The pK_a Values of Acidic and Basic Residues Buried at the Same Internal Location in a Protein Are Governed by Different Factors. Journal of Molecular Biology. 2009; 389(1):34–47. [PubMed: 19324049]

21. Isom DG, Cannon BR, Castaneda CA, Robinson A, Bertrand GME. High tolerance for ionizable residues in the hydrophobic interior of proteins. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105(46):17784–17788. [PubMed: 19004768]

22. Isom DG, Castaneda CA, Cannon PD, Velu PD, Garcia-Moreno B. Charges in the hydrophobic interior of proteins. Proceedings of the National Academy of Sciences of the United States of America. 2010; 107(37):16096–16100. [PubMed: 20798341]

23. Warshel A. Calculations of Enzymatic Reactions: Calculations of pK$_a$, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. Biochemistry. 1981; 20(11):3167–3177. [PubMed: 7248277]

24. Mehler EL, Fuxreiter M, Simon I, Garcia-Moreno EB. The Role of Hydrophobic Microenvironment in Modulating pK_a Shifts in Proteins. Proteins: Stru Func Genet. 2002; 48(1): 283.

25. Rekker, RF. The Hydrophobic Fragmental Constant. Nauta, WT.; Rekker, RF., editors. Amsterdam: Elsevier; 1977.

26. Rekker RF. The Hydrophobic Fragmental Constant; an Extension to a 1000 Data Point Set. Eur J Med Chem. 1979; 14:479–488.

27. Rekker, RF.; Mannhold, R. Calculation of Drug Lipophilicity: The Hydrophobic Fragmental Constant Approach. Weinheim: VCH; 1992.

28. MacKerell AD Jr, Bashford D, Bellott M, Dunbrack RL Jr, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE III, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B. 1998; 102(18):3586–3616.

29. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. Journal of Chemical Physics. 1996; 105(5):1902–1921.

30. Wisz MS, Hellinga HW. An Empirical Model for Electrostatic Interactions in Proteins Incorporating Multiple Geometry-Dependent Dielectric Constants. PROTEINS. 2003; 51(3):360–377. [PubMed: 12696048]

31. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: New developments. Nucleic Acids Res. 2008; 36(Database issue):D419–425. [PubMed: 18000004]

32. Nguyen DM, Leila Reynald R, Gittis AG, Lattman EE. X-ray and Thermodynamic Studies of Staphylococcal Nuclease Variants I92E and I92K: Insights into Polarity of the Protein Interior. Journal of Molecular Biology. 2004; 341(2):565–574. [PubMed: 15276844]

33. Hooft RW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins. 1996; 26(4):363–376. [PubMed: 8990493]

34. Nielsen JE, Vriend G. Optimizing the Hydrogen-Bond Network in Poisson-Boltznann Equation-Based pK_a Calculations. Proteins. 2001; 43(4):403–412. [PubMed: 11340657]

35. Nielsen JE, McCammon JA. On the evaluation and optimization of protein X-ray structures for pKa calculations. Protein Science. 2003; 12(2):313–326. [PubMed: 12538895]

36. Spassov VZ, Yan L. A fast and accurate computational approach to protein ionization. Protein Science. 2008; 17(11):1955–1970. [PubMed: 18714088]

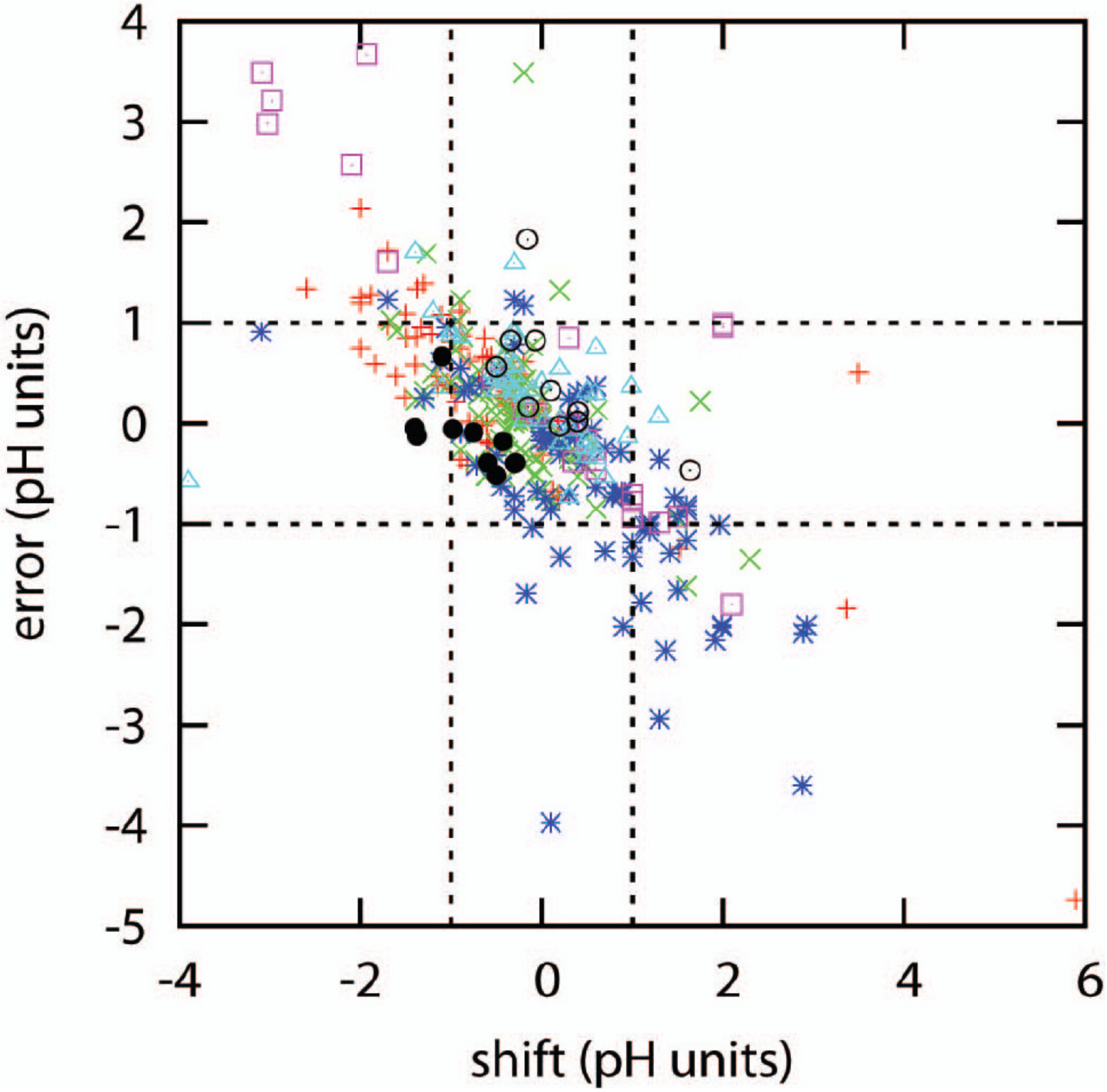

Figure 1.

pK_a prediction errors vs. pK_a shifts of 334 TR in 54 proteins. Data for different residue types are in different colors and symbols. Asp is in red "+", Glu in green "x", His in blue stars, Tyr in purple open squares, Lys in cyan triangles, N- and C-terminal residues in black circles and dots, respectively. Dotted horizontal and vertical lines are for error = ±1 pH unit and shift = ±1 pH unit, respectively.

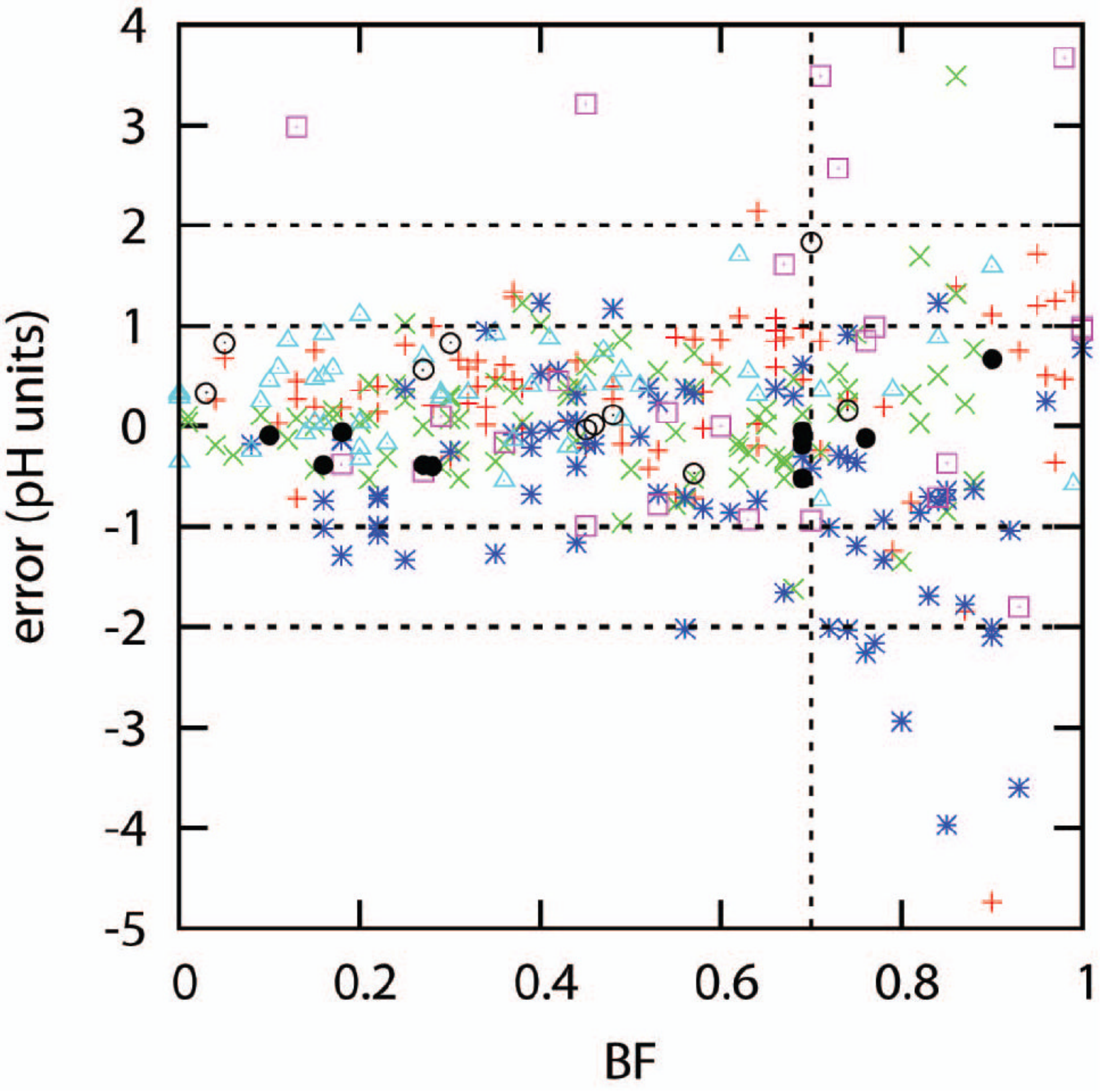

Figure 2.
pK_a prediction errors vs. BF of 334 TR in 54 proteins. Colors and symbols, see Figure 1. Dotted horizontal lines are for error = ±1, ±2 pH units and vertical line is for BF = 0.7.

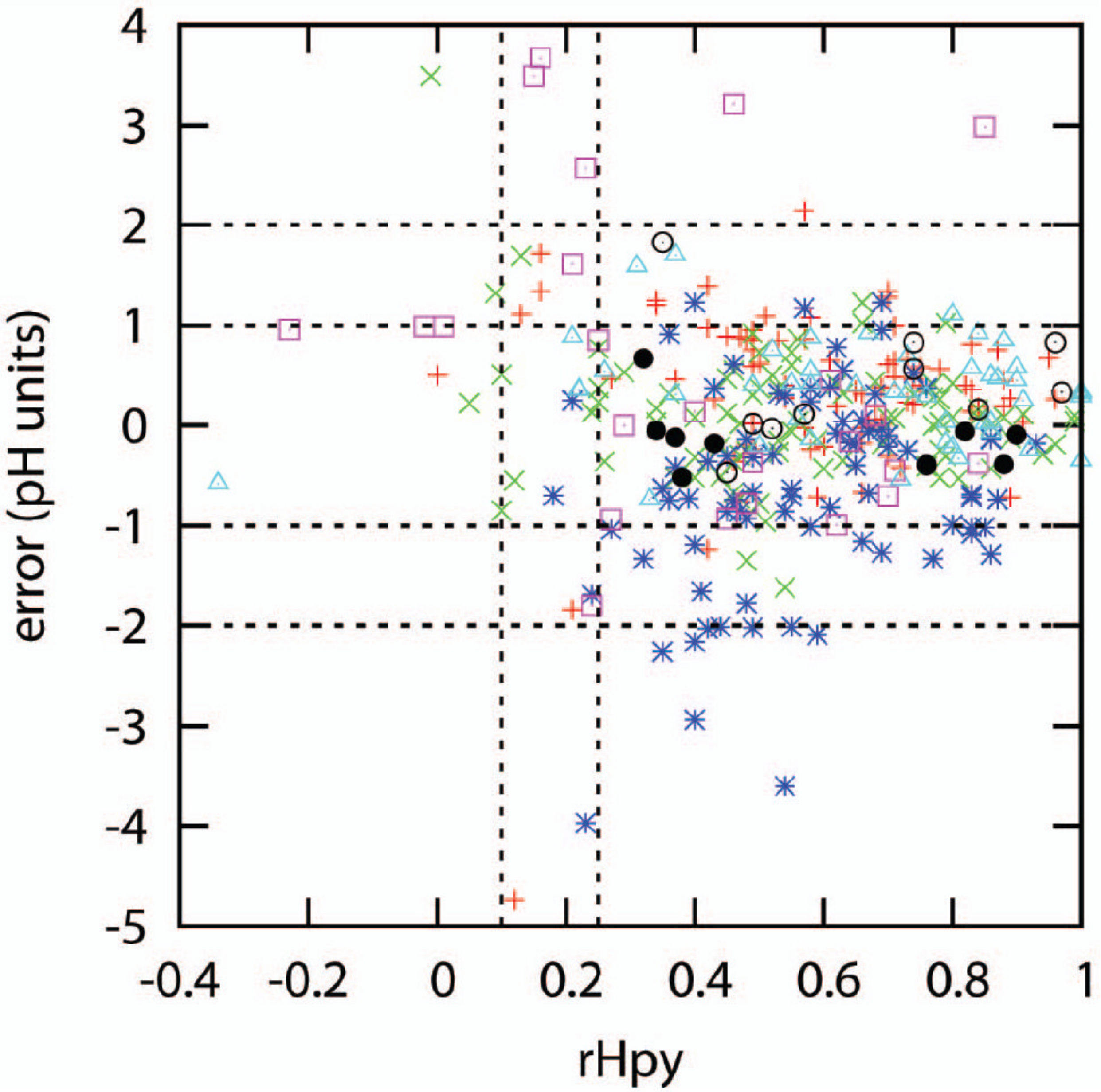

Figure 3.

pK_a prediction errors vs. rHpy of 334 TR in 54 proteins. Colors and symbols as in Figure 1. Dotted horizontal lines are for error = ±1, ±2 pH units and vertical lines are for rHpy = 0.1 and 0.25, respectively.

Table I

Calculated pK_a of SNase mutants with known experimental pK_a.

	protein[a]	mutant	pK_a[b]	exp[c]	error	BF[d]	rHpy[e]	α[f]
1	2oxp	V66D	6.8	8.1	-1.3	0.968	0.02	3.
2	2oxp(m)	V66D	8.2	8.1	-0.1	0.972	-0.13	3.
3	1tqo	I92E	8.6	9.0	-0.4	0.985	-0.42	3.
4	1tr5(RT)	I92E	8.7	9.0	-0.3	1.00	-0.35	3.
5	1tt2	I92K	5.7	5.3	0.4	0.998	-0.49	3.
6	2rks	L38K	9.9	>10.4	-0.5	0.929	0.20	1.
7	2rks(m)	L38K	9.5	>10.4	-0.9	0.926	0.20	1.
8	3d6c	L38E	5.3	6.8	-1.5	0.995	0.14	1.
9	3d6c(m)	L38E	6.8	6.8	0.0	0.998	0.07	2
10	3ero(A)	I72E	5.9	7.3	-1.4	0.858	0.32	1
11	3ero(B)	I72E	8.6	7.3	1.3	0.919	-0.09	3

[a] RT = room temp., m= locally minimized structure, all mutants not specifically marked (RT) were done at cryogenic temperatures.

[b] Calculated pK_a.

[c] Experimental pK_a (Errors for threshold values are calculated from pK_a – threshold value).

[d] Buried fraction.

[e] Normalized hydrophobicity, see text following Eq. (5).

[f] Transfer energy scaling factor (see text).

Table II

Calculated pK_a of SNase mutants with unknown experimental pK_a values.[a]

pdb code	mutant	pK_a	exp	err	BF	rHpy	α
3d4w	A109R	10.9	>10.4	0.5	0.890	0.27	1
3dhq	A90R	10.7	>10.4	0.3	0.930	0.16	1
2rbm	I72K	11.2	8.6	2.6	0.582	0.58	1
3erq	L25K	5.8	6.2	-0.4	1.00	-0.44	3
3evq	L25E	8.7	7.5	1.2	0.995	-0.31	3
3eji	L36K	7.8	7.2	0.6	1.00	-0.11	3
3dmu	T62K	7.3	8.1	-0.8	0.998	0.01	3
3d4d	Y91E	7.3	7.1	0.2	0.977	0.03	2
3e5s	L103K	9.8	8.2	1.6	0.994	0.19	1
3clf	V104K	8.7	7.7	1.0	1.00	0.09	2
3c1e	L125K	5.9	6.2	-0.3	1.00	-0.08	3
rmsd				1.1 (2.9)[b]		1	

[a] See footnotes, Table I, for column header definitions.

[b] Value in parentheses is rmsd from the null hypothesis.

Table III

pK_a prediction accuracy for 334 TR in 54 proteins.

# of residues		% of residues predicted with error		
		<1 unit	< 1.5 unit	< 2 unit
Total [a]	334	82	91	94
Shift < 1unit [b]	237	94	98	99
Shift ≥ 1unit [c]	97	53	74	81

[a] Total number of TR in the extended data set.

[b] Number of TR with experimental pK_a shifted by < 1 pH unit.

[c] Number of TR with experimental pK_a shifted by ≥ 1 pH unit.

Table IV

RMSD of pK_a prediction errors for each protein SCOP class.

SCOP Class [a]	RSMD (# of residues)		
	total	shift < 1 unit	shift ≥ 1 unit
a	0.90 (73)	0.87 (55)	0.99 (18)
b	0.89 (52)	0.78 (36)	1.10 (16)
c	1.11 (70)	0.46 (46)	2.33 (24)
d	0.71 (100)	0.54 (69)	0.99 (31)
g	0.36 (32)	0.30 (39)	0.71 (4)

[a] SCOP Class: *a*, all α proteins; *b*, all β proteins; *c*, α and β proteins (*a*/*b*); *d*, α and β proteins (*a*+*b*); *g*, small proteins.

Table V

RMSD of pK_a prediction errors for each residue type.

Residue Type	RSMD (No. of residues)		
	Total [a]	shift < 1 unit [b]	shift ≥ 1 unit [c]
Asp	0.92 (79)	0.48 (48)	1.34 (31)
Glu	0.67 (85)	0.61 (74)	0.99 (11)
His	1.18 (80)	0.88 (51)	1.58 (29)
Tyr	1.65 (24)	0.40 (9)	2.07 (15)
Lys	0.61 (46)	0.52 (39)	0.94 (7)
N-Terminus	0.73 (10)	0.76 (9)	0.47 (1)
C-Terminus	0.35 (10)	0.42 (7)	0.39 (3)

[a] RSMD for all residues in certain residue type.

[b] RMSD for certain residues with pK_a shifted by < 1 pH unit.

[c] RMSD for certain residues with pK_a shifted by ≥ 1 pH unit.