PKA17 – a coarse-grain grid-based methodology and web-based software for predicting protein pKa shifts

John P. Cvitkovic, Connor D. Pauplis, George A. Kaminski*

Department of Chemistry and Biochemistry, Worcester Polytechnic Institute, 100 Institute Rd., Worcester, MA 01609

KEYWORDS Protein pKa shifts, Coarse-grain models

ABSTRACT: We have developed and tested PKA17, a coarse-grain grid-based model for predicting protein pKa shifts. Our pKa predictor is currently deployed via a website interface. We have carried out parameter fitting using 442 Asp, Glu, His, Lys, and Arg residues for which experimental results are available in the literature. PROPKA software has been used for benchmarking. The average unsigned error and RMSD have been found to be 0.628 and 0.831 pH units, respectively, for PKA17. The corresponding results with PROPKA are 0.761 and 1.063 units. We have assessed the robustness of the developed PKA17 methodology with a number of tests and have also explored the possibility of using a combination of PROPKA and PKA17 calculations in order to improve the accuracy of predicted pKa values for protein residues. We have also once again confirmed that protein acidity constants are influenced almost entirely by residues in the immediate spacial proximity of the ionizable amino acids. The resulting PKA17 software has been deployed online with a web-based interface at http://users.wpi.edu/~jpcvitkovic/pka_calc.html

I. Introduction

Assessing protein acidity constant values is important in predicting the structure, stability, and function of proteins. Thus, it is beneficial to be able to predict the values of these acidity constants computationally when robust experimental data are not available.

Protein pKa values have to be calculated in aqueous solutions to be biochemically relevant. These acidity constants are proportional to the total free energies of deprotonation. Such a deprotonation free energy is a sum of the bond breaking energy and the free energy of hydration for the resulting ions. The two components have opposite signs and large magnitudes, often hundreds of kcal/mol. The accuracy of calculating the final pKa value thus depends on reproducing or predicting a very fine balance of energies, since computational predictions need to be accurate within ca. 0.8-1.0 pH unit or slightly over 1 kcal/mol in order to be relevant. This is why robust calculations of protein and other acidity constants remain a very difficult task even in the presence of the computational resources available today.

Thus, a number of research groups have applied significant efforts to achieve reliable and accurate results in the computational assessment of protein pKa values. Protein pKa values are proportional to the free energies of deprotonation in aqueous solution.It is convenient to calculate pKa shifts instead of the absolute values, the former being the differences introduced into the acidity constants (and thus deprotonation free energies) by transferring the ionizable group into the protein environment from the hydrated form of a simple reference compound. For example, propanoic acid can be used as the reference compound for the aspartic acid residue. Thus, the task is to assess the difference in the deprotonation energy due to the interactions of the residue with the other parts of the protein as opposed to its interactions with bulk solvent. The goal has been seen as predicting protein pKa values within 0.8 – 1.0 pH units.[1]

The efforts in computational prediction of protein pKa shifts have been made in several directions. It is natural to apply the Poisson-Boltzmann equation to the calculation of acidity constants,[2,3] and a number of variations and approximations of this general methodology for pKa calculations have been suggested.[4-19] Other techniques have been applied as well.[1] The Poisson-Boltzmann (PB) approach can be implemented in volume- or surface-based formalisms.[1] A variety of approaches have been proposed to optimize the electrostatic charges used in PB simulations and to address the need to take into the account the presence of multiple ionizable residues.[9,10] Some research groups

have suggested assigning a dielectric constant with large magnitude (up to 20) to the interior of proteins.[11-13] While this approach led to an overall improvement of the results, it still left a number of calculated pKa values deviating significantly from experimental values.[1,14] Furthermore, the very physical meaning of such high values of the dielectric constant for the protein interior is not clear.

It has also been acknowledged that conformational changes in the protein in response to the protonation or deprotonation of ionizable residues have to be taken into account to improve the accuracy of protein acidity constant predictions. Techniques involving ensembles of conformers (usually with side-chain variations) have been proposed.[20-28] Of particular interest is the Multi-Conformation Continuum Electrostatic (MCCE) method that combines motion of side-chains with continuum dielectric treatment of solvent and bulk protein.[29-33] It has been used in many applications, including successfully predicting pKa values for an extensive testing set of several hundred protein residues with AMBER, CHARMM, and PARSE force fields. The prediction results were compared with the experimentally measured acidity constants.[33]

A number of microscopic techniques with an explicit treatment of solvent have also been suggested.[34-38] Some of these techniques use quantum mechanical representation of the systems.[39-42] While quantum methodology is generally more accurate and rather potent, it also requires greater amounts of computational resources than non-quantum empirical techniques, and thus its use is currently somewhat limited when protein pKa calculations are to be carried out.[1]

Combined quantum mechanical/molecular mechanical (QM/MM) methods can offer a better alternative than purely quantum simulations.[39-42] Some techniques also employ complete or partial continuum representation of solvent. One successful example is in applications of constant pH molecular dynamics (CPHMD) simulations.[43-46] In many cases, constant pH simulations approach or match first-principles level of accuracy .[1] Additionally, CPHMD techniques offer a tool for studying pH-dependent conformational phenomena.

At the same time, much attention has been directed toward development and application of empirical techniques of evaluating protein pKa shifts, with PROPKA being one of the most successful and widely used examples.[47-49] In these cases, some physical considerations are combined with statistical fitting of descriptors and parameters that predict amino acid pKa values depending on the environment of the particular ionizable residue. Statistical fitting methods can yield a reasonably high level of accuracy and such methods implicitly replace any conformational and rotamer sampling that may be needed to account for thermal motion. Such methods are sufficiently accurate in most cases, and they are also very fast and robust when applied to diverse sets of protein residues.

We have created PKA17, a predictor of protein pKa shifts that has been parameterized on a subset of experimentally known acidity constants of protein residues. One of its distinguishing features is that it employs an extremely coarse-grain model of the protein with each residue represented as only a single particle. This makes the model very simple and at the same time reduces the noise levels, as fine variations in atomic positions have no effect on the calculated values of the acidity constants. The other distinguishing feature is the use of a cubic grid model for positioning of the protein residues. Finally, the physical formalism of PKA17 is much simpler than that of PROPKA, with the fitting being almost entirely statistical with care taken to only introduce a minimum number of fitable parameters. At the same time, we believe that this statistics is defined by the underlying physical principles. We have used 442 protein residues for fitting and benchmarking of our model. In spite of its simplicity, PKA17 was found to perform on par with – or slightly better than –PROPKA. We have also tested an approach in which PROPKA and PKA17 are used in a combination to predict protein pKa shifts, and we found this combined approach is capable of giving better results than either of the techniques alone.

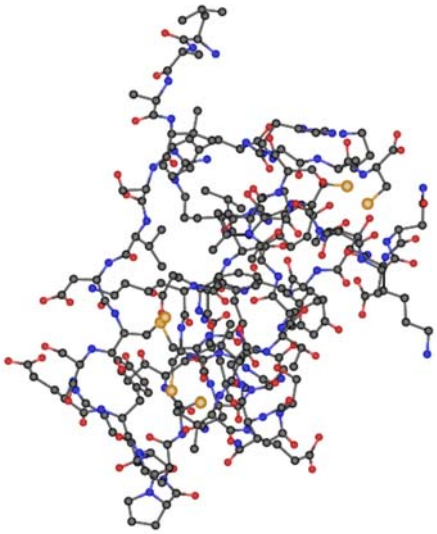
The rest of the paper is organized as follows. Methods are presented in Section II. Given in Section III are results and discussion. Summarizing conclusions can be found in Section IV.

II. Methods

Mapping of the Protein Geometry to the Grid

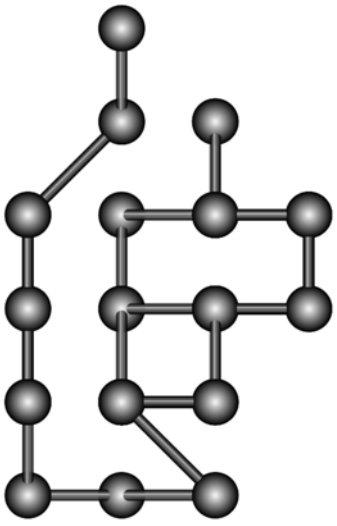
Each protein residue is represented by a single particle the location of which is determined solely by the position of the alpha carbon of the residue. The Cartesian coordinates of the alpha carbons are taken from the input PDB file and then mapped onto nodes of a cubic grid. The side of each grid cell is set to be 5.4 Å. This makes each cubic cell to have a volume that is approximately equal to the average volume of a protein residue.

The process of the geometry mapping by PKA17 is illustrated in Figure 1 using chain I from the 1ppf PDB structure. We start with the full PDB structure (a). Then we parse it to leave only the alpha carbon locations, each of those representing the whole residue (b). Finally, each of such particles is placed at the nearest node of the cubic lattice with 5.4 Å spacing in each dimension (c). The types of the residues and the connections to the adjacent residues are recorded and retained at this stage. The grid-mapped structure along with the residue type and connectivity information is passed to the next stage of the pKa calculation as described in the following subsection.



(a)

(b)



(c)

Figure 1. Schematic depiction of the process of mapping protein residue coordinates onto the cubic grid. The mapping proceeds from the full atomistic PDB structure (a) to the locations of the alpha carbons (b) and finally to the cubic grid nodes (c).

Determining Residues that Define the Values of the pKa Shifts

The current version of PKA17 predicts acidity constants for five types of protein residues – Asp, Glu, His, Lys, Arg. Each of these types has an initial reference pKa constant of A_i^0, where the subscript denotes the amino acid type. It should be noted that this constant is not intended to correspond to the pKa value for the residue in any particular protein or peptide, as it is always modified by influence from other residues. The final value of the acidity constant is determined by the following sum:

$$pK_a = A_i^0 + \sum B_{ij} + \sum C_{ij} + \sum D_{ij} + \sum E_{ij} \quad (1)$$

The meanings of the terms of Eq. 1 are illustrated in Figure 2. Coefficient B_{ij} signifies the shift in the pKa value of residue *A* of type *i* resulting from being directly connected to a residue of type *j* in the backbone. Coefficient C_{ij} stands for the effect of on the acidity constant of residue *A* of type *i* induced by a non-connected residue of type *j* located just one lattice period *l* away. Coefficient D_{ij} represents the effect of a diagonally placed residue of type *j*, and E_{ij} shows the influence of a residue of type *j* placed at a distance of $\sqrt{3} \cdot l$. Residues that are located farther away do not affect the pKa value of residue *A*.

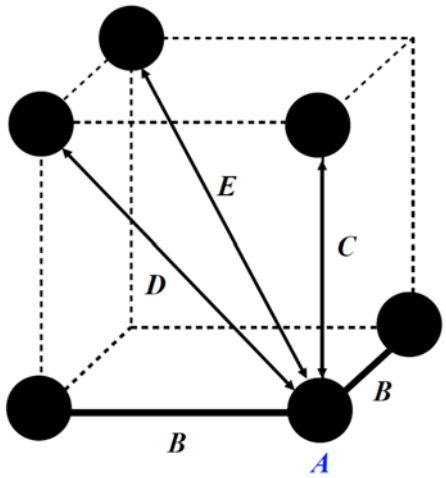


Figure 2. Neighboring residues affecting pKa value of residue *A*.

The values of the coefficients that determine the pKa value of residue *A* of type *i* depend on the type, *j*, of the influencing residue. Non-connected residues that are mapped to the same node of the cubic grid as residue *A* are assigned the same pKa shift coefficients, B_{ij}, as those residues within exactly distance *l* (the grid spacing) from it.

The above mapping and coefficients are all that determine the values of protein pKa shifts in the PKA17 framework. Values of all the parameters are found by fitting to experimentally measured values, with further tests on proteins and amino acids that were not included in the fitting set.

It should be noted that we are not utilizing any explicit procedure for establishing whether a residue is exposed to the solvent or buried within a protein. However, the shift of the acidity constant does depend on the number of neighboring residues, and thus the effect of exposure to or separation from the solvent is automatically included in an implicit way. Additionally, while the desolvation contribution to the pKa shift cannot be separated from the other factors contributing to the pKa shift, it is accounted for as a part of the *B, C, D*, and *E* coefficients in Equation 1.

III. Results and Discussion

Target Data Sets Used in Fitting of the pKa Shift Coefficients

In the present version of the PKA17 software, $B_{ij} = C_{ij} = D_{ij} = E_{ij}$ for any *i* and *j*. In other words, any residue of type *j* shifts the pKa value of residue of type *i* by the same amount if the residues are no farther than $\sqrt{3} \cdot l$ apart. It also does not matter whether the residues are covalently bonded, only the geometric distance between the grid-mapped alpha carbons is used to determine if the residue pairs are neighbors. This was done to avoid overparameterization and related issues with stability and transferability of the results. Naturally, the coefficients are different for different pairs of residue types *i* and *j*. Moreover, it should be emphasized that, in general $B_{ij} \neq B_{ji}$ and the same is true for coefficients *C*, *D*, and *E*.

We used an extensive fitting and testing set of protein residue pKa values from Reference 50. A complete list of the proteins and residues can be found in the Supplementary material file.

The general fitting procedure was as follows. For each of the ionizable residue types that we considered (Asp, Glu, His, Lys), we divided the set of experimental pKa values from the literature into two subsets. The first part was the fitting set. We fitted the parameters A_i^0 and $B_{ij} = C_{ij} = D_{ij} = E_{ij}$ for this residue type to minimize the deviation of the acidity constants calculated with the PKA17 software from the experimental results. The resulting average deviation constituted the first benchmarking result for our fitting.

Then we obtained the leave-one-out (LOO) average unsigned errors. In this case, one residue was excluded from the fitting, and the resulting fitted parameters were employed to calculate the pKa value for this residue. The procedure was repeated for all the residues in the set. Therefore, we essentially assessed the average errors for residues that were not a part of the fitting procedure at all.

The next step was in applying the parameters derived for the full fitting set to calculate pKa values of the test set. This was done without any refitting.

Finally, we used the full set (fitting and testing sets together) to fit the final set of the PKA17 parameters and to calculate the LOO average unsigned error. While the LOO result was used as a benchmarking measure, the final parameter set that is currently used in the software is the one obtained by fitting to the full set of residues.

All the calculated errors were compared with those produced with the PROPKA website,[51] as PROPKA is one of the most successful and widely used web-based protein pKa predictors.

The following subsections contain results for the specific residue types.

Fitting and Testing Results for Aspartic Acid Residues

We used a fitting set of 105 residues and a testing set of 33 residues (the complete list of the proteins, residues, and the results is given in the Supplementary Materials file). The results are summarized in Table 1 and Figure 3.

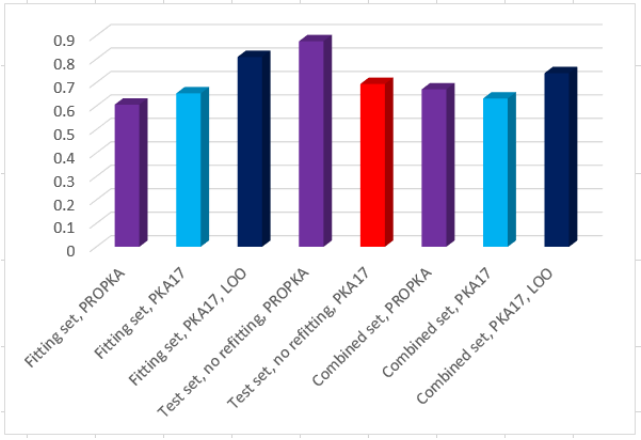


Figure 3. Average unsigned errors in pKa values of Asp residues.

Fitting PKA17 parameters for Asp to the fitting set values resulted to an averaged unsigned error of 0.654 pH units, which is a bit higher that the PROPKA 0.606 pH units. The leave-one-out (LOO) average error was 0.809 pH units. While this number is somewhat higher than the PROPKA errors, it should be noted that the PROPKA training set does include some of the residues that were employed in our tests.

The performance of this intermediate set of PKA17 parameters for the test subset of the Asp residues was better than that of the PROPKA software, with the average errors being 0.876 and 0.694 pH units for PROPKA and PKA17, respectively.

Finally, the fitting to the complete combined set of Asp residues lead to an average unsigned error of 0.632 pH units as calculated with PKA17. The LOO average error was 0.740 pH units. The PROPKA result is 0.671 pH units. Once

again, we need to keep in mind that some of the residues were used in parameter fitting for PROPKA.

Fitting and Testing Results for Glutamic Acid Residues

In this case, the fitting and the testing sets consisted of 101 and 32 residues, respectively. The results are summarized in Table 2 and Figure 4.

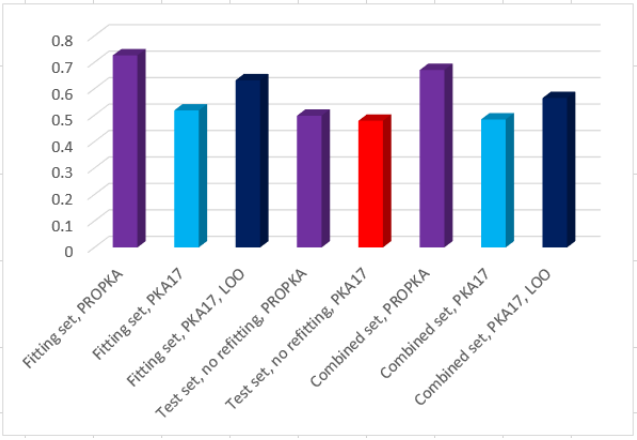


Figure 4. Average unsigned errors in pKa values of Glu residues.

Both the fitted PKA17 parameters for the fitting set and the LOO calculations yield lower average unsigned errors of 0.518 and 0.632 pH units, respectively, than the average PROPKA error of 0.726 units. The non-fitted test set results are somewhat similar (0.479 pH units error for PKA17 and 0.498 units for PROPKA). Using the complete set that includes both the fitting and the testing subsets for the fitting leads to an average PKA17 error of 0.484 and the LOO average unsigned error of 0.565 pH units, both numbers being lower than the average PROPKA error of 0671 units.

It should be noted explicitly that one of the major reasons for including the leave-one-out errors is the need to test the stability of the resulting PKA17 framework with respect to the fitting data set and its ability to predict acidity constants for residues that are not a part of the fitting set at all. At the same time, it makes sense to minimize the actual final error and to employ the fitting that employs every single residue to produce the finalized version of the PKA17 parameters.

Fitting and Testing Results for Histidine Residues

We used a fitting subset of 61 His residues and a testing subset of 28 residues, comprising a full combined set of 89 histidine residues. The results of our fitting and testing for histidine are summarized in Table 3 and Figure 5.

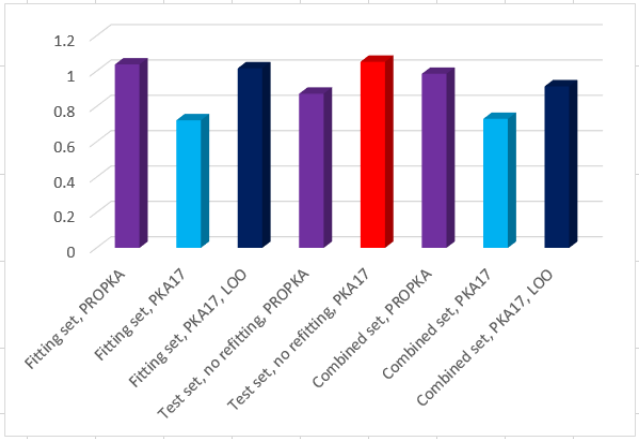


Figure 5. Average unsigned errors in pKa values of His residues.

PKA17 performs better than PROPKA for this residue type. For the fitting set, the fitting of the parameters for the PKA17 framework permits to lower the average unsigned error to 0.722 pH units, while the PROPKA result is a ca. 43% higher error of 1.038 units. Even with the leave-one-out (LOO) approach, the PKA17 error is only 1.016 pH units. When applied to the test set, the PKA17 error is somewhat higher (1.053 units) than the PROPKA one (0.872 units). However, using the complete combined set, we obtained 0.730 units and 0.914 units average errors with the fitting and LOO PKA17 runs, while the PROPKA average unsigned error is 0.985 pH units. Overall, the PKA17 performance in these histidine calculations is sufficiently robust.

Fitting and Testing Results for Lysine Residues

The composition of our lysine fitting and testing sets was somewhat different than that of the Asp, Glu, and His ones. Most of Lys residues have pKa values within a relatively narrow range. Our main fitting set is composed of such cases. The testing set (that was also a part of the complete combined set) included mutants that exhibit a greater variety of pKa values.[50e] This way we covered a broader range of potential lysine acidity constant shits that can be encountered in wild or engineered proteins.

The results are summarized in Table 4 and Figure 6.

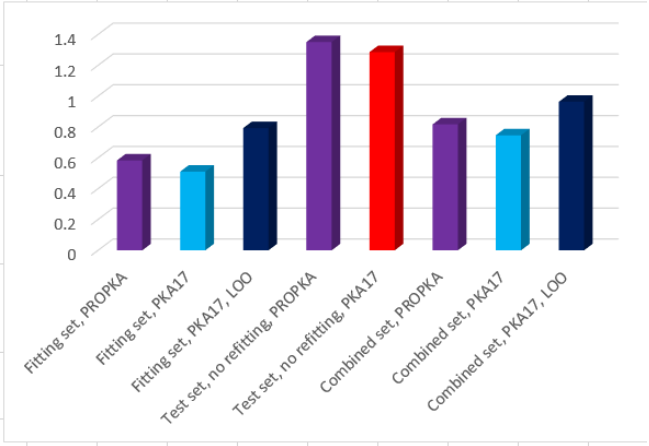


Figure 6. Average unsigned errors in pKa values of Lys residues.

The initial fitting set contained 57 Lys residues. The average unsigned error in the pKa values obtained wit PROPKA was 0.583 pH units, and the PKA17 average error with fitting for this set was 0.510 units. The LOO PKA17 result was 0.793 pH units. It should be emphasized again that it is hard to make a direct comparison between PROPKA and PKA17 given that PROPKA parameters were developed with some training/fitting on this particular data set as well.

Calculating Lys pKa values for the 25 residue test set (composed of residues with a greater range of pKa shifts than the fitting set) yields average PROPKA and PKA17 errors of 1.351 and 1.286 pH units, respectively. These results are fairly similar, but PKA17 seems to perform somewhat better for these structures that were not employed in the direct initial fitting. When we use the complete combined set that contains all the 82 residues, the average error with PROPKA is 0.817 pH units, while the fitted and leave-one-our average errors produced with PKA17 are 0.746 and 0.964 units, respectively.

Using the PROPKA Fitting Sets for Asp and Glu pKa values

The fitting sets used to train PROPKA performance for the aspartic and glutamic acid residues are available from the literature.[48] We have employed these sets in order to produce a more direct comparison of the PKA17 and PROPKA results.

The Asp PROPKA fitting set contains 43 residues (the full list is given in the Supplementary Materials). The average unsigned error for the pKa values for the set computed with PROPKA is 0.503 pH units. When we use the same set for fitting PKA17 parameters, the average error is only 0.299 pH units. The leave-one-out (LOO) procedure resulted in an average error of 0.460, which is still 8.5% lower than the average PROPKA deviation.

We then applied the resulting PKA17 parameters to calculating pKa values of all the 138 Asp residues in our complete combined aspartic acid set. The average PROPKA error (also reported in the corresponding subsection above) was 0.671 pH units, while the PKA17 one was slightly higher at 0.768 units. It is worth recalling the average errors produced by the PKA17 calculations after fitting to the complete combined Asp set. The average error in complete fitting was 0.632 pH units, while the LOO procedure lead to an error of 0.740 units, which is rather close to the 0.768 units resulting from fitting to the PROPKA training set. While both the complete fitting and the LOO results obtained with the direct fitting to the combined set are better, the LOO one is rather close to the result obtained with the parameters produced in fitting to the PROPKA training set. We believe that this indicates the robustness of the PKA17 framework with respect to the fitting protocol and fitting set of residues. The results of the Asp fitting tests are summarized in Table 5 and Figure 7.

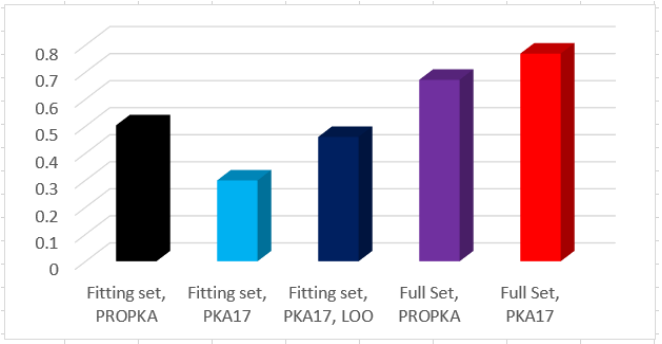


Figure 7. Average unsigned errors in pKa values of Asp residues after fitting to the PROPKA training set.

The PROPKA fitting set for glutamic acid contains 42 residues.[15] The average unsigned error for these residues that we obtained with PROPKA is 0.469 pH units. Fitting for to this set of resides for PKA17 results in an average error of 0.331 units, while the LOO protocol gives an error of 0.471 units, which is virtually the same as the PROPKA one.

Application of the resulting parameters to the full combined set for Glu that contains 133 residues yields the following results. The average unsigned PROPKA error in pKa values is 0.671 pH units. The error obtained with PKA17 is 0.603 units, or about 10% lower.

The results of using the PROPKA training set for the Glu residue are summarized in Table 6 and Figure 8.

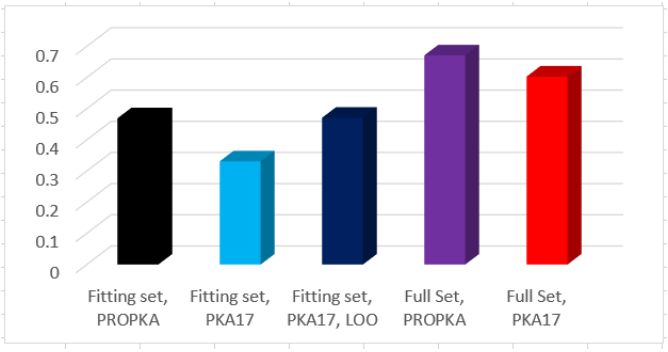


Figure 8. Average unsigned errors in pKa values of Glu residues after fitting to the PROPKA training set.

Overall, we can conclude from the above results that the PKA17 fitting methodology is robust and stable with respect to the choice of the fitting set. The resulting PKA17 parameters permit results that are at least as good as the PROPKA ones when both programs are trained and tested on the same data sets.

Comparative Timing of PROPKA and PKA17

While neither of the programs takes prohibitively long to produce results, we still ran a brief comparison of the required computational time for five representative proteins. The results are shown in Table 7 and Figures 9 and 10.

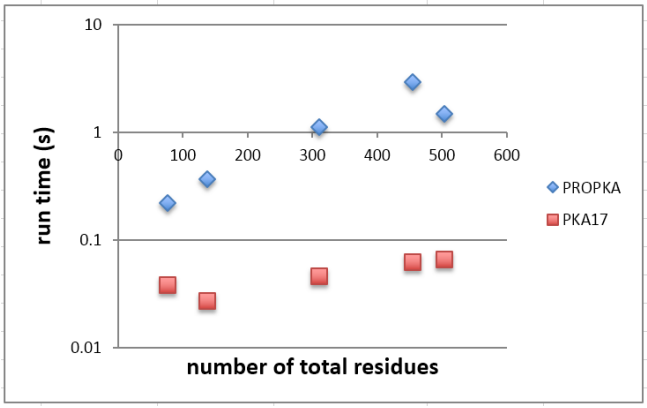


Figure 9. Computational time required by PROPKA and PKA17 software as a function of the total number of residues in the protein.

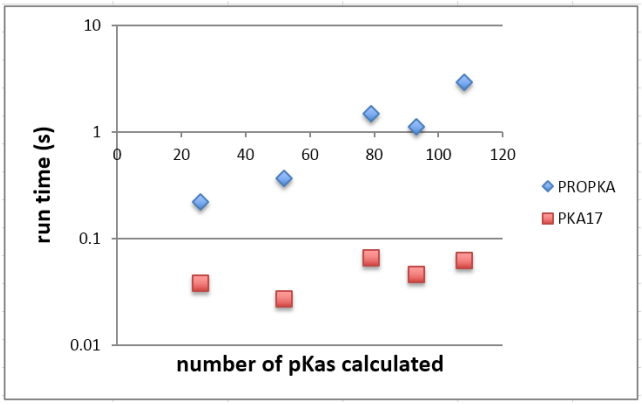


Figure 10. Computational time required by PROPKA and PKA17 software as a function of the number of pKa values calculated for the protein.

Both programs are rather fast, but it is still worth noting that PKA17 is an order of magnitude faster than PROPKA. The very simple model we are using for mapping the proteins and determining the pKa shifts is actually both fast and robust.

Using PROPKA and PKA17 Together

We have also calculated pKa values for the complete combined sets for Asp, Glu, His, and Lys by combining the PROPKA and PKA17 results for each residue in equal proportion (0.5 of PROPKA pKa + 0.5 of PKA17 pKa). Resulting average unsigned errors are shown in Table 8 and Figure 11.

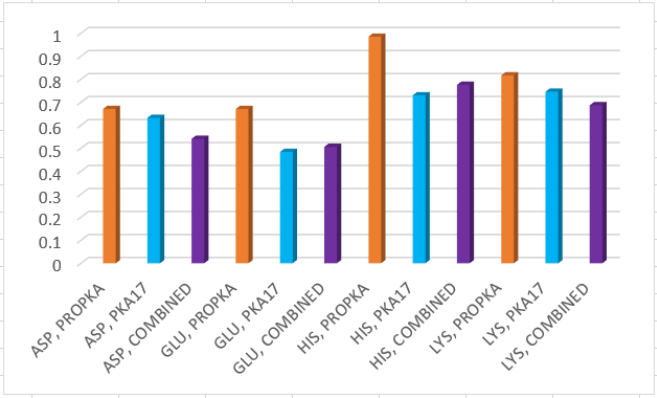


Figure 11. Average unsigned errors in pKa values calculated with PROPKA, PKA17, and as the average of the PROPKA and PKA17 values.

It can be seen that this combined application offers a clear advantage. All the combined PROPKA/PKA17 average errors are lower than those resulting from applying PROPKA alone. The errors for Asp and Lys are also lower than those for applying PKA17 alone. Apparently, PROPKA and PKA17 tend to err in the opposite directions, and their combination provides a more robust option for predicting protein pKa values.

Benchmarking PKA17 Using an Extensive MCCE Test Set

The Multi-Conformation Continuum Electrostatic (MCCE) method using a Poisson-Boltzmann approach with AMBER, CHARMM, and PARSE force fields has been developed and successfully applied by Alexov and coworkers in order to calculate pKa values of an extensive set of protein residues.[33] We have used the same dataset to provide additional validation of our technique by calculating pKa values with the PKA17 and PROPKA software.

The comparison of accuracy in calculating these acidity constants for Asp, Glu, His, and Lys residues are given in Tables 9-12 and on Figures 12-15. The overall accuracy is presented in Table 13 and on Figure 16. All the pKa values for the individual residues are listed in the supporting information file. It should be emphasized that these results were obtained with no additional fitting of the PKA17 parameters.

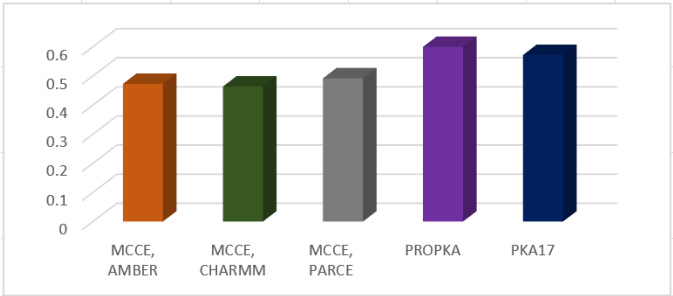


Figure 12. Average errors of Asp pKa calculations for the extensive fitting set presented in Reference 33.

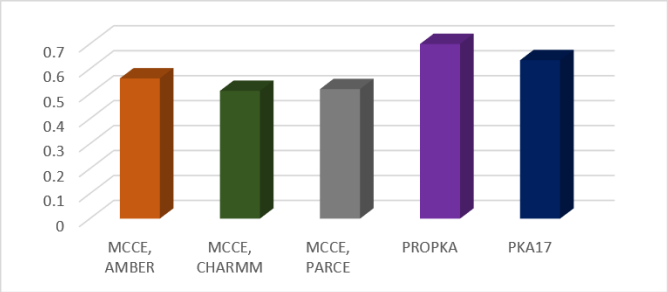


Figure 13. Average errors of Glu pKa calculations for the extensive fitting set presented in Reference 33.

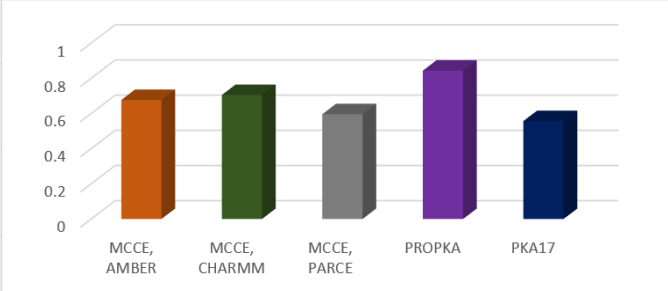


Figure 14. Average errors of His pKa calculations for the extensive fitting set presented in Reference 33.

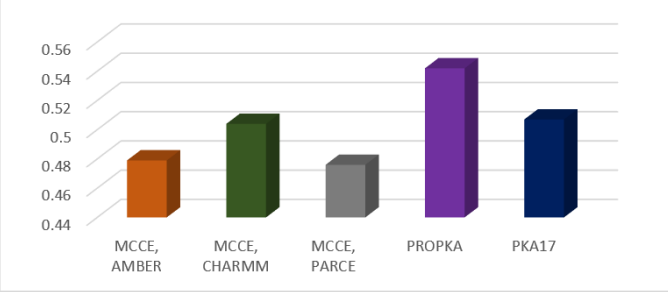


Figure 15. Average errors of Lys pKa calculations for the extensive fitting set presented in Reference 33.

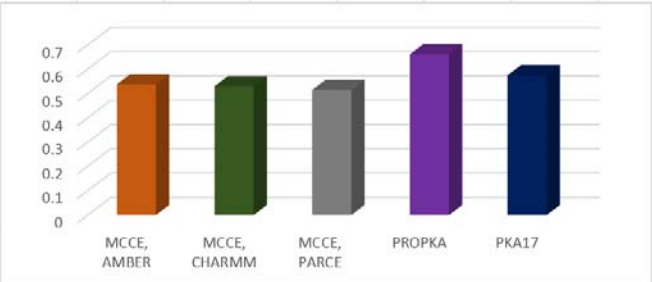


Figure 16. Overall average unsigned errors of pKa calculations for the extensive fitting set presented in Reference 33.

It can be seen from these tables and figures that, in almost all of the cases (including the overall results), the PKA17 formalism performs somewhat worse than the much more sophisticated MCCE technique, but slightly better than the PROPKA suite. Histidine calculations are an exception, with the PKA17 results being better than those produced by all the other techniques.

We believe that these tests confirm the suitability of the PKA17 formalism and implementation for use in the general field of fast prediction of protein pKa shifts.

IV. Conclusions

We have developed and validated a predictor of protein pKa values named PKA17. It is based on a coarse-grain grid model of proteins. The pKa shifts are defined by the residues that are spatially close to the ionizable residue in question. Fitting and validation of our model involved an extensive set of 441 protein residues. Additional benchmarking on a previously proposed extensive set of ionizable protein residues[33] appears to confirm the robustness of the presented technique. The resulting tool has been deployed with a web-based interface at http://users.wpi.edu/~jpcvitkovic/pka_calc.html

The results demonstrate that PKA17 performs on par or even somewhat better than the widely used and successful protein pKa predictor PROPKA. It also requires less computational resources, with the computational time needed for PKA17 runs being an order of magnitude lower than that required by PROPKA. Moreover, we have achieved the current level of accuracy with PKA17 while significantly limiting the number of fitting variables in order to avoid any danger of possible overparameterization.

We have also shown that the accuracy of PKA17 is reasonably robust with respect to the choice of the fitting set for parameterization, even though some more sophisticated techniques (such and MCCE and quantum mechanics) can yield a higher degree of accuracy of evaluating pKa shifts of protein residues.

In addition to the above, we have tested a combined application of PKA17 and PROPKA and found that such a combination results in improved accuracy of the pKa predictions compared to either technique used individually.

Therefore, we report creation of an accurate and efficient web-interfaced protein pKa predictor, and, we have demonstrated that protein pKa shifts can be assessed with a set of very simplistic coarse-grid predictors. Moreover, all these predictors are local, with only the immediate environment of the ionizable residue affecting the final acidity constant value.

ASSOCIATED CONTENT

Supporting Information. All the experimental and calculate pKa values used as in the fitting and testing data sets.

Funding Sources

This project was partially supported by the NSF REU grant CHE 1659529.

AUTHOR INFORMATION

Corresponding Author

*George A. Kaminski, Department of Chemistry and Biochemistry, Worcester Polytechnic Institute, 100 Institute Rd, Worcester, MA 01609. Email: gkaminski@wpi.edu.

REFERENCES

(1) Alexov, E.; Mehler, E. L.; Baker, N.; Baptista, A. M.; Huang, Y.; Milletti, F.; Nielsen, J. E.; Farrell, D.; Carstensen, T.; Olsson, M. H. M.; Shen, J. K.; Warwicker, J.; Williams, S.; Word, J. M. *Proteins* **2011**, *79*, 3260–3275.

(2) Tanford C.; Kirkwood J. G. *J Am Chem Soc* **1957**; *79*: 5333–5339.

(3) Tanford C.; Roxby R. *Biochemistry* **1972**; *11*: 2192–2198

(4) Bashford D,; Karplus M. *Biochemistry* **1990**; *29*: 10219–10225.

(5) Potter M,; Gilson M,; McCammon J. *J Am Chem Soc* **1994**; *116*: 10298–10299.

(6) Nielsen J,; McCammon A. *Protein Sci* **2003**; *12*: 313–326.

(7) Dolinsky T. J.; Nielsen J. E.; McCammon J. A.; Baker N.A. *Nucleic Acids Res* **2004**; *32*: W665–W667.

(8) Yang A-S.; Gunner M. R.; Sampogna R.; Sharp K.; Honig B. *Proteins* **1993**; *15*: 252–265.

(9) Gilson M. K. *Proteins* **1993**; *15*: 266–282.

(10) Lim C.; Bashford D.; Karplus M. *J Phys Chem* **1991**; *95*: 5610–5620.

(11) Antosiewicz J.; McCammon J.; Gilson M. *J Mol Biol* **1994**; *238*: 415–436.

(12) Antosiewicz J.; Briggs J.; Elcock A.; Gilson M.; McCammon J. *J Comp Chem* **1996**; *17*: 1633–1644.

(13) Antosiewicz J. M.; Cammon J. A.; Gilson M. K.. *Biochemistry* **1996**; *35*: 7819–7833.

(14) Teixeira V. H.; Cunha C. A.; Machuqueiro M.; Oliveira A. S.; Victor B. L.; Soares C. M.; Baptista A. M. *J Phys Chem B* **2005**; *109*: 14691–14706.

(15) Karshikoff A. *Protein Eng* **1995**; *8*: 243–248.

(16) Baptista A.; Soares C. *J Phys Chem B* **2001**; *105*: 293–309.

(17) Warwicker J. *Protein Sci* **2004**; *13*: 2793–2805.

(18) Nielsen J.; Andersen K.; Honig B.; Hooft R.; Klebe G.; Vriend G.; Wade R. *Protein Eng* **1999**; *12*: 657–662.

(19) Nielsen J.; Vriend G. *Proteins* **2001**; *43*: 403–412.

(20) You T.; Bashford D. *Biophys J* **1995**; *69*: 1721–1733.

(21) Beroza P,Case D. *J Phys Chem* **1996**; *100*: 20156–20163.

(22) Kieseritzky G.; Knapp E. W. *Proteins* **2008**; *71*: 1335–1348.

(23) Barth P.; Alber T.; Harbury P. B. *Proc Natl Acad Sci USA* **2007**; *104*: 4898–4903.

(24) Warwicker J.; Watson H. C. *J Mol Biol* **1982**; *157*: 671.

(25) Warwicker J. *J Theor Biol* **1986**; *121*: 199–210.

(26) Warwicker J. *Protein Sci* **1999**; *8*: 418–425.

(27) Koehl P.; Delarue M. *J Mol Biol* **1994**; *239*: 249–275.

(28) Cole C.; Warwicker J. *Protein Sci* **2002**; *11*: 2860–2870.

(29) Alexov E.; Gunner M. *Biophys J* **1997**; *74*: 2075–2093.

(30) Georgescu R.; Alexov E.; Gunner M. *Biophys J* **2002**; *83*: 1731–1748.

(31) Georgescu R.; Alexov E.; Gunner M. *Biophys J* **2002**; *83*: 1731–1748.

(32) Song Y.; Mao J.; Gunner M. R.; *Comput Chem* **2009**; *30*: 2231–2247.

(33) Wang, L.; Li, L.; Alexov, E. *Proteins* **2015**; *83*: 2186–2197.

(34) Warshel A.; Russell S. *Q Rev Biophys* **1984**; *17*: 283–422

(35) Warshel A. Computer modeling of chemical reactions in enzymes and solutions. New York: John-Wiley & Sons, Inc.; 1991.

(36) Mehler E. L. *Theor Comput Chem* **1996**; *3*: 371–405.

(37) Schulz C.; Warshel A. *Proteins* **2001**; *44*: 400–417.

(38) Warshel A.; Levitt M. *J Mol Biol* **1976**; *103*: 227–249.

(39) Li H.; Robertson A. D.; Jensen J. H. *Proteins* **2004**; *55*: 689–704.

(40) Jensen J. H.; Li H.; Robertson A. D.; Molina P. A. *J Phys Chem A* **2005**; *109*: 6634–6643.

(41) Li H.; Robertson A.; Jensen J. **2004**; *55*: 689–704.

(42) Schaefer P.; Riccardi D.; Cui Q. *J Chem Phys* **2005**; *123*: 014905.

(43) Khandogin J.; Brooks C. L., III. *Biochemistry* **2006**; *45*: 9363–9373.

(44) Lee M. S.; Salsbury F. R., Jr., Brooks C. L., III. *Proteins* **2004**; *56*: 738–752.

(45) Khandogin J.; Brooks C. L.,III. *Biophys J* **2005**; *89*: 141–157.

(46) Wallace J. A.; Shen J. K. *J Chem Theory Comput* **2011**; *7:* 2617–2629.

(47) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704-721.

(48) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 525-537.

(49) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 2284-2295.

(50) (a) Castaneda, C. A.; Fitch, C. A.; Majumdar, A.; Khangulov, V.; Schlessman, J. L.; Garcia-Moreno, B. E. *Proteins* **2009**, *77*, 570-588; (b) Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. *Protein Sci.* **2009**, *18*, 247-251; (c) Arthur, E. J.; Yesselman, J. D.; Brooks, C. L. *Proteins* **2011**, *79*, 3276-3286; (d) Hiebler, K.; Lengyel, Z.; Castaneda, C. A.; Makhlynets, O. V. *Proteins* **2017**, *85*, 1656-1665; (e) Wu, X.; Lee, J. Brooks, B. R. *J. Phys. Chem. B*, **2017**, *121*, 3318-3330;

(51) http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/, accessed last on June 25, 2018.

TABLES.

Table 1. Results of pKa calculations for Asp residues (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	0.606
Fitting set, PKA17	0.654
Fitting set, PKA17, LOO	0.809
Test set, PROPKA	0.876
Test set, PKA17	0.694
Combined set, PROPKA	0.671
Combined set, PKA17	0.632
Combined set, PKA17, LOO	0.740

Table 2. Results of pKa calculations for Glu residues (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	0.726
Fitting set, PKA17	0.518
Fitting set, PKA17, LOO	0.632
Test set, PROPKA	0.498
Test set, PKA17	0.479
Combined set, PROPKA	0.671
Combined set, PKA17	0.484
Combined set, PKA17, LOO	0.565

Table 3. Results of pKa calculations for His residues (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	1.038
Fitting set, PKA17	0.722
Fitting set, PKA17, LOO	1.016
Test set, PROPKA	0.872
Test set, PKA17	1.053
Combined set, PROPKA	0.985
Combined set, PKA17	0.730
Combined set, PKA17, LOO	0.914

Table 4. Results of pKa calculations for Lys residues (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	0.583
Fitting set, PKA17	0.510
Fitting set, PKA17, LOO	0.793
Test set, PROPKA	1.351
Test set, PKA17	1.286
Combined set, PROPKA	0.817
Combined set, PKA17	0.746
Combined set, PKA17, LOO	0.964

Table 5. Results of Aspartic acid pKa calculations after fitting to the PROPKA training set (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	0.503
Fitting set, PKA17	0.299
Fitting set, PKA17, LOO	0.460
Combined set, PROPKA	0.671
Combined set, PKA17	0.768

Table 6. Results of Glutamic acid pKa calculations after fitting to the PROPKA training set (in pH units)

Method	Average unsigned error in pKa values
Fitting set, PROPKA	0.469
Fitting set, PKA17	0.331
Fitting set, PKA17, LOO	0.471
Combined set, PROPKA	0.671
Combined set, PKA17	0.603

Table 7. CPU time required for pKa calculation (in seconds)

PDB ID	1ubq	2dhc	2gga	3twy	4pyp
Number of residues	76	310	455	137	504
Number of pKa values	26	93	108	52	79
Time, PROPKA	0.219	1.121	2.947	0.369	1.489
Time, PKA17	0.038	0.046	0.062	0.027	0.066

Table 8. Average unsigned errors in pKa values calculated with PROPKA, PKA17, and as the average of the PROPKA and PKA17 values (in pH units)

System/Method	Average unsigned error in pKa values
Asp	
PROPKA	0.671
PKA17	0.632
Combined/Averaged	0.541
Glu	
PROPKA	0.671
PKA17	0.484
Combined/Averaged	0.506
His	
PROPKA	0.985
PKA17	0.730
Combined/Averaged	0.776
Lys	
PROPKA	0.817
PKA17	0.746
Combined/Averaged	0.687

Table 9. Average errors of Aspartic acid pKa calculations for the extensive fitting set presented in Reference 33. (in pH units)

Method	Average unsigned error in pKa values
MCCE, AMBER[33]	0.473
MCCE, CHARMM[33]	0.464
MCCE, PARCE[33]	0.492
PROPKA	0.600
PKA17	0.572

Table 10. Average errors of Glutamic acid pKa calculations for the extensive fitting set presented in Reference 33. (in pH units)

Method	Average unsigned error in pKa values
MCCE, AMBER[33]	0.562
MCCE, CHARMM[33]	0.512
MCCE, PARCE[33]	0.519
PROPKA	0.700
PKA17	0.635

Table 11. Average errors of Histidine pKa calculations for the extensive fitting set presented in Reference 33. (in pH units)

Method	Average unsigned error in pKa values
MCCE, AMBER[33]	0.677
MCCE, CHARMM[33]	0.705
MCCE, PARCE[33]	0.596
PROPKA	0.844
PKA17	0.557

Table 12. Average errors of Lysine pKa calculations for the extensive fitting set presented in Reference 33. (in pH units)

Method	Average unsigned error in pKa values
MCCE, AMBER[33]	0.479
MCCE, CHARMM[33]	0.504

MCCE, PARCE[33]	0.476
PROPKA	0.542
PKA17	0.507

Table 13. Overall average errors of pKa calculations for the extensive fitting set presented in Reference 33. (in pH units)

Method	Average unsigned error in pKa values
MCCE, AMBER[33]	0.538
MCCE, CHARMM[33]	0.530
MCCE, PARCE[33]	0.515
PROPKA	0.661
PKA17	0.575

GRAPHICAL ABSTRACT