12-2011

Progress in the Prediction of pKa Values in Proteins

Emil Alexov
Clemson University, ealexov@clemson.edu

Ernest L. Mehler
Cornell University

Nathan Baker
Pacific Northwest National Laboratory

Antonio Baptista
Universidade Nova de Lisboa, Portugal

Yong Huang
Washington University in St Louis

See next page for additional authors

Authors
Emil Alexov, Ernest L. Mehler, Nathan Baker, Antonio Baptista, Yong Huang, Francesca Milletti, Jens Erik Nielsen, Damien Farrell, Tommy Carstensen, Mats H.M. Olsson, Jana K. Shen, Jim Warwicker, Sarah Williams, and J Michael Word

Published in final edited form as:
Proteins. 2011 December ; 79(12): 3260–3275. doi:10.1002/prot.23189.

PROGRESS IN THE PREDICTION OF pK_a VALUES IN PROTEINS

Emil Alexov[1], **Ernest L Mehler**[2], **Nathan Baker**[3], **Antonio Baptista**[4], **Yong Huang**[5], **Francesca Milletti**[6], **Jens Erik Nielsen**[7], **Damien Farrell**[8], **Tommy Carstensen**[8], **Mats H. M. Olsson**[9], **Jana K. Shen**[10], **Jim Warwicker**[11], **Sarah Williams**[12], and **J. Michael Word**[13]

[1]Department of Physics, Clemson University, Clemson, USA [2]Physiology and Biophysics, Weill Medical College of Cornel University, USA [3]Pacific Northwest National Laboratory, USA [4]Instituto de Tecnologia Química e Biológica, Portugal [5]Dept. of Biochemistry and Molecular Biophysics, Washington University in St. Louis, USA [6]University Studi Perugia, Italy [7]University College Dublin, Dublin, Ireland [8]School of Biomolecular and Biomedical Science, Ireland [9]Department of Chemistry, University of Copenhagen, Denmark [10]Department of Chemistry and Biochemistry, University of Oklahoma, USA [11]Faculty of Life Sciences, University of Manchester, UK [12]Chemistry & Biochemistry, University of California at San Diego, USA [13]OpenEye Scientific Software, Inc., USA

Abstract

The pK_a-cooperative aims to provide a forum for experimental and theoretical researchers interested in protein pK_a values and protein electrostatics in general. The first round of the pK_a-cooperative, which challenged computational labs to carry out blind predictions against pK_as experimentally determined in the laboratory of Bertrand Garcia-Moreno, was completed and results discussed at the Telluride meeting (July 6–10, 2009). This paper serves as an introduction to the reports submitted by the blind prediction participants that will be published in a special issue of *PROTEINS: Structure, Function and Bioinformatics*. Here we briefly outline existing approaches for pK_a calculations, emphasizing methods that were used by the participants in calculating the blind pK_a values in the first round of the cooperative. We then point out some of the difficulties encountered by the participating groups in making their blind predictions, and finally try to provide some insights for future developments aimed at improving the accuracy of pK_a calculations.

Keywords

pKa; protein electrostatics; pH dependent properties of proteins; predicting pKa values in proteins

STATEMENT OF PURPOSE OF THE pK_a-COOPERATIVE

Computational and experimental study of acid-base equilibria in proteins has reached a point where further progress in increasing the reliability of predicting pK_a's will require the development of new approaches that better describe the underlying physics regulating the system's structure and dynamics as well as any pH-dependent phenomena [1]. Such improvements may be based on entirely novel algorithms or on combining the strongest components of existing approaches. To carry out the latter, an initial step will be the detailed analysis of the strengths and weaknesses of existing approaches. Toward that end participants in a workshop on protein electrostatics, organized by Marilyn Gunner and

Correspondence to: Emil Alexov; Ernest L Mehler.

Bertrand Garcia-Moreno, concluded that it was timely to assess the different methods for calculating pK_a, how they would fare on some difficult cases, and subsequently how these approaches could be improved. It was decided that the best framework for accomplishing this goal was to establish a (preliminary) cooperative that would be a repository of data and act as a channel for bringing together researchers who are active in developing and applying methods for calculating acid/base dissociation constants in proteins. The first meeting of the pK_a-cooperative was held at Telluride, July 6–10, 2009. This paper is a summary of that meeting.

To provide a focus for the meeting, research groups involved in pK_a calculations were asked to make blind predictions using the extensive structural and experimental results on Staphylococcus nuclease (SNase) provided by the Garcia-Moreno group. This group had determined structures and measured various p*K*a of wild type SNase and a large number of mutants [2–10]. The results of the blind predictions were discussed at the meeting, and thanks to the willingness of all contributors to discuss their results in an open forum, the meeting was successful in identifying a number of issues relevant to improving the accuracy of pKa prediction. The open discussion allowed the group to avoid the fatal pitfall for this type of exercise to degenerate into a competition with “winners” and “losers”. The avoidance of such a trap is essential if the entire community is to profit from comparing the different methods and gain insight into how to incorporate improvements. The usefulness of making blind predictions is their objectivity for testing a given method because of the impossibility of “improving” the results by further refinement of the parameters. Thus blind predictions provide a measure of the state of development of a particular approach and gives clues where improvements are to be made. This paper serves as an introduction to the special issue of *PROTEINS: Structure, Function and Bioinformatics* that will report the results from the individual groups that participated in the blind prediction exercise.

In the next section, we give a brief overview of methods used in pK_a calculations, but concentrating on the methods used by the participants of the meeting, and then a section that is based on the experiences of the blind contributors. We asked them to write a short description of their calculations, but without including any results. We were particularly interested in problems and difficulties that were encountered during the calculations. Finally, in a concluding section, we briefly consider future directions and speculate (“predict”) on how to develop methods that are both accurate and not too computationally demanding. The ultimate goal is to not only predict pK_as but to reveal the underlying physics regulating the ionization.

OVERVIEW OF METHODS FOR CALCULATING pK_as IN PROTEINS

Introduction

The calculation of the pK_a of titratable groups in proteins had its beginning in the work of Tanford and Kirkwood based on the Poisson-Boltzmann equation (BPE) [11]. This early work provided methods for studying acid-base equilibria in proteins even before the 3-dimensional structure of any protein was known. With the development of x-ray crystallography as a powerful tool for the accurate determination of protein structure and the introduction of computers, it became possible to calculate the pK_as of titratable groups in proteins at ever-increasing levels of detail and complexity. In particular, with the significant increase in computing power over the last decade, there has been a rapid development of novel methods for calculating pK_as that, in principle, are able to give an accounting of the underlying physics that controls the acid-base equilibrium of the titrating systems in a protein. At the time that the initial use of the PBE as a tool for calculating pKa was being explored, physical chemists turned to the evaluation of dissociation constants in bifunctional acids and bases. Their approach was to express the electrostatic free energy of interaction of

the bifunctional groups by $\Delta w = q_1 q_2 / D_e R$ where R is the distance between the charges, q, and D_e is an effective screening [12]. As with the PBE, the so-called screened Coulomb potential (SCP) has been the starting point of many modern methods for calculating pK_a in proteins.

Unfortunately, the reliability of calculated pK_as has not kept pace with the development of new and more sophisticated methods for modeling titratable systems: errors of two or more pH units in calculated pK_a values are not unusual. In particular, errors of over 1 pK_a unit are most likely in predicted values for titratable residues where the measured pK_a indicates a large shift from the reference value. Such errors are particularly troublesome for cases where residue pK_a values shift into the physiological pH range. Errors in calculated pK_a values for highly-perturbed residues are a serious issue because many studies report pK_a calculations on a subset of the titratable residues in one or a few proteins and, if the results are satisfactory, conclude that the method works. Experience suggests, however, that the reliability of a given method can only be assessed after applying the approach to many proteins of different structural characteristics [13]. A mitigating factor in some cases is that absolute accuracy in the pK_a value is not essential for rationalizing pH dependent processes in biological macromolecules, where the protonation state of key titratable residues at physiological pH, or changes in pK_a with structural transitions is often sufficient to develop useful insights into the physical mechanism of a biological process.

The development of experimental methods to determine pK_a values has also seen rapid progress and the introduction of NMR techniques [14–17] has made pK_a measurements accurate and fairly routine in globular proteins. Thus a large (and still growing) body of data is now available that can be used to test the computational approaches. Some experimentalists have developed and made available systematic data sets of values consisting of wild type and mutant proteins that can be used to carefully probe the computational methods to identify the sources of disagreement between calculated and experimental results. Such probing will hopefully lead to improvement of the computational methods.

Most methods for predicting pK_a values in proteins are based on estimating the additional free energy terms that appear when the protonatable moiety is transferred from solvent into the protein, which formally can be expressed as:

$$pK_a\,(\text{protein}) = pK_a\,(\text{solvent}) + \Delta pK_a\,(\text{solvent} \Rightarrow \text{protein}) \quad (1)$$

The first term on the right hand side provides a reference value representing the pK_a of the residue in the solvent (typically termed the null model), while the second term comprises all the new interactions that arise from removing the residue from the pure solvent (desolvate) and embed it in the protein (resolvate), which itself is immersed in the solvent. Theoretical and experimental evidence indicates that the most important class of interactions that determine ΔpK_a are electrostatic in origin. Therefore, to be able to predict pK_a values reliably, a reasonably accurate description of the electrostatics and other relevant energy terms in the protein and surrounding environment is required. Minimally, the description of the electrostatics must comprise a term that describes the Coulomb interactions between the charges that model the protein structure and a term that describes the interaction of the charges with the solvent, often termed the “self” or “transfer” energy. The importance of the latter term was pointed out long ago by Warshel [18]. It is noted that recent simulation results suggest that inclusion of other components of the intermolecular potential, e.g., hydrophobic effects, may also improve the predictions. [19]

The calculation of the electrostatic effects can be based on a microscopic or macroscopic framework. Truly macroscopic models express the system by a continuum description and assume that the required quantities can be calculated directly from the macroscopic electrostatics equations; *i.e.*, the PBE. Microscopic models calculate all interactions at the atomic level of detail, and thermodynamic properties are obtained by statistical averaging. There is broad agreement that ultimately it is most desirable to use the microscopic framework because of its greater theoretical content. However, many microscopic methods tend to be computationally expensive and therefore, in most cases, macroscopic continuum approaches have been used because of these computational limitations. Fortunately, this issue is gradually being resolved by the availability of ever increasing computing power and more efficient methods for simulation and sampling. As a result, there has been an increasing interest in developing microscopic methods.

In recent years, a new class of methods has been developed that is based on the large data set of measured protein pK_a values that is now available. These new methods are purely empirical in concept and use the protein structure to account for the different types of interactions; *e.g.*, H-bonds or charge-charge interactions, and assign each such interaction an energetic weight that is optimized using the large data base of experimentally determined pK_a values or titration curves. The advantage of these methods is their speed, but their disadvantage is that they are not physics-based and thus provide less physical insight into the determinants of shifted pK_a values. The results of these methods seem quite reasonable provided one is within the radius of convergence defined by the data set used in the parameterization, but extrapolation is likely to be less satisfactory until the data base is extended. This implies, e.g., that the effects of mutation on the pKa of a particular group may be in error even though the wild type (WT) pKa is correctly predicted. In contrast, even though methods based on the electrostatic equations often require empirical parameters to yield reasonable results, it may still be possible to rationalize the underlying physics that leads to the shifted pKa.

Below, we briefly review pK_a prediction methods starting with macroscopic approaches followed by microscopic approaches and finally empirical methods (see for example Ref. [20]). It is noted that this review is not meant to be exhaustive, but primarily concentrates on the methods that were of interest and discussed at the 2009 Telluride meeting.

Macroscopic methods

Although most physics based methods for calculating pK_as are based on either “macroscopic” or “microscopic” models, some formulations are mixed, juxtaposing macroscopic and microscopic quantities. A typical example is using molecular dynamics (MD) with an implicit solvent description such as Generalized Born (GB). Resolving this juxtaposing of mutually inconsistent quantities in a physically reasonable way may be part of the difficulty experienced in formulating reliable methods for calculating pH dependent quantities.

(a) PB equation based methods—The earliest methods for calculating pK_a values represented the protein by an impenetrable sphere because the resulting PBE could be solved analytically. The most influential of these methods was developed by Tanford and Kirkwood (TK) [11] and Tanford and Roxby[21], based on a model where the protein was represented by an impenetrable sphere of radius *b* with embedded titratable points and a low dielectric constant, and an exterior region with a high dielectric constant. The TK method was introduced before any protein structures had been solved, but as soon coordinates became available the TK method was modified to account for the solvent accessibility of the titratable group since it was argued that charges near the protein surface would experience

additional damping due to the polar solvent [22,23]. Subsequently, many other modifications have been proposed [24]; nevertheless, the advent of large scale computing machines has allowed the use of numerical methods which can solve the PBE directly for proteins of any shape.

Proteins and other biological macromolecules are irregularly shaped multi-atomic objects existing in water in the presence of mobile ions. The electrostatic potential (φ) in such a system can be calculated using the PBE, i.e.,

$$\nabla \cdot \varepsilon(\mathbf{r})\nabla\varphi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) + \varepsilon(\mathbf{r})\kappa^2(\mathbf{r})\sinh(\varphi/k_B T) \quad (2)$$

where $\varepsilon(\boldsymbol{r})$ is the dielectric permittivity, ρ(r) is the permanent charge density, κ is the Debye-Huckel parameter, k_B is Boltzmann constant and T is temperature.

For irregularly shaped objects the PBE does not have analytical solutions, so that, the electrostatic component of the solvation energy and the corresponding ion screening must, in practice, be calculated with numerical solutions, of which several approaches are available. The most frequently used numerical methods of solving the PBE can be grouped into two distinct categories: methods implemented on volume-filling grids (including finite difference, finite volume, and finite element methods) and boundary element (BE) methods where the solution is expressed in terms of distributions over the molecular surface. Commonly used PB solvers include (1) DelPhi developed in the Honig lab [25–27], (2) APBS developed by Baker and coworkers [28,29] and several new additions made in the McCammon lab [30–32], (3) CHARMM [33] is a molecular mechanics and simulation program that includes a FD based PB solver developed by Roux and co-workers [34], (4) ZAP developed by Nicholls and co-workers [35], (5) MEAD developed by Bashford [36], (6) AFMPB solver developed by Lu and co-workers [30], and (7) MIBPB developed by Wei and co-workers [37].

Bashford and Karplus pioneered the field of PB-based methods for predicting pK_as of ionizable groups. They developed a macroscopic electrostatic continuum model using detailed structural information to treat self-energies and interactions arising from permanent partial charges and titratable charges [38] and solved the PBE using finite difference methods. Testing the approach on lysozyme resulted in the observation that the pK_a values are very sensitive to the details of the local protein conformation, and that side-chain mobility is likely to be important in determining the observed pK_a shifts. It is also of note that the accuracy of the pK_a values already hinted at the issues that would develop around the definition of the dielectric constant.

The PB-based approach was also used by McCammon and co-workers [39,40] to predict pK_a values using 3D structures of the corresponding proteins/small molecules. Wade and co-workers showed that the optimization of the parameters such as partial charges could significantly improve the pK_a predictions [41]. The Baker and Nielsen groups collaborated successfully to develop a set of tools for pK_a calculations [42]. Honig and co-workers further improved the FDPB method for calculating pK_as [43]. The novelty of their technique with respect to previous work was the specific incorporation within the numerical protocol of both the neutral and charged forms of each ionizable group. The multiple-site titration algorithm [44] developed by Gilson and co-workers addressed the necessity of computing pK_as of proteins having large number of titratable sites, resulting in an exponentially-growing number of possible charged or uncharged states. Based on the results in Ref. [45] a pragmatic approach was taken by Antosiewicz and co-workers to account for conformational flexibility through the use of a high dielectric constant of 20 for the protein

interior [46–48]. This procedure seemed to improve overall results, but left several important titration sites in serious error. Baptista and co-workers[49] investigated the use of two distinct protein dielectric constants for computing the individual (site) and the pairwise (site-site) terms of the ionization free energies, but they found no overall improvement over the use of a single value of 20, even for buried or shifted sites. Karshikoff further explored the use of the dielectric constant to mimic protein flexibility [50] by assigning different local dielectric constants per residue type with a combination of the FDPB and Tanford-Roxby iterative procedures. In addition, Baptista and coworkers proposed the methodology of computing pK_as with alternative hydrogen positions [51]. The method of Warwicker and co-workers [52] estimated the conformational relaxation in a pH-titration with a mean-field assessment of maximal side chain solvent accessibility. Another FDPB-based method was introduced by Nielsen and co-workers, which adds an explicit step to optimize the hydrogen bonds network. It was shown that this approach delivers better results than methods not optimizing the hydrogen bond network [53,54].

(b) The PBE and conformational flexibility—It became evident that protein conformational flexibility should be explicitly taken into consideration within the same protocol that calculates pK_as. Bashford and co-workers introduced polar proton conformational flexibility into the pK_a protocol [55] by generating an ensemble of conformers where the positions of polar protons were systematically varied. This information was then used to explicitly calculate intrinsic pK_a values and electrostatic interactions between titrating sites. The method was applied to the Asp, Glu, and Tyr residues of hen lysozyme. Different protocols for hydrogen atom placement were used and their effect tested against experimental pK_a values. It was determined that multi-conformational calculations significantly improved the agreement with experiment. The subsequent Monte-Carlo based method of Beroza and Case [56] included side chain flexibility in continuum electrostatic calculations of protein titration. Knapp and co-workers [57] demonstrated that the geometry and the hydrogen bonding are very important in treating pK_as of residues involved in salt bridges. Hartbury and co-workers recently developed a rotamer repacking method called FDPB-MF that exhaustively samples side chain conformational space and rigorously calculates multibody protein-solvent interactions [58]. Their method achieved high accuracy on a small subset of acidic residues in turkey ovomucoid third domain, hen lysozyme, Bacillus circulans xylanase, and human and Escherichia coli thioredoxins, with root mean square deviations of 0.3 pH units [58]. Recently, Warwicker and coworkers developed the FD/DH method [52], which is an automated combination of Finite Difference Poisson-Boltzmann (FDPB) [59,60] and Debye-Hückel (DH) methods. This is based on the well-known finding that ΔpK_as for water accessible groups are generally dominated by water dielectric, and can be handled in a simple DH model with relative dielectric of 78.4, whereas solvent exclusion can lead to larger ΔpK_as, handled better by FDPB with separate water and protein dielectrics [61]. The code statistically averages pK_as over multiple conformers and multiple FDPB calculations. In the FD/DH method, a short-cut approximation avoids multi-conformation sampling, with DH interactions only being sampled where assessment of maximal solvent accessible surface area (SASA) for an ionisable group is greater than a fixed fraction [52]. This assessment is made with a mean-field sampling of side chain rotamer packing on a fixed backbone. [62,63]

One of the most commonly used method for incorporating conformational flexibility into pK_a calculations combines FDPB electrostatic calculations with explicit sampling of side chain, hydrogen and ligand positions. This approach, developed by Gunner and co-workers, is known as the Multi-Conformation Continuum Electrostatics (MCCE) method[64–67]. In the MCCE the protein side chain motions are simulated explicitly while the dielectric effect of solvent and bulk protein material is modeled by continuum electrostatics. MCCE can be used to: (1) study the protein structural responses to changes in charge; (2) study the changes

in charge state of ionizable residues due to structural changes in the protein; (3) study the structural and ionization changes caused by changes in solution pH; (4) find the location and stoichiometry of proton transfers coupled to electron transfer; (5) make side chain rotamer packing predictions as a function of pH. Recently Alexov and co-workers developed a hybrid pK_a method that uses distinctive different ensembles of structures representing conformational ensemble for ionized and neutral forms of the titratable residue of interest. These ensembles were generated either with MD simulations or *ab-initio* structure predictions. Then the structures were subjected to MCCE calculations and the pK_as were predicted by averaging the corresponding titration curves.

(c) Generalized Born—As an alternative to PBE, a computationally faster approach based on Born's theory of ionic solvation was developed. This approach is based on an early extension of the Born formula (proposed by Hoijtink [68] to allow the Born approach to be applied to systems with a distribution of N point charges and was expressed in the form

$$\Delta G^{pol} = \frac{1}{2}(\varepsilon_s^{-1} - 1)\sum_{i,j}^{N} \frac{q_i q_j}{r_{ij} + \delta_{ij} R_i} \tag{3}$$

where q_i is the net charge (not necessarily integral) on particle i, r_{ij} is the separation between q_i and q_j, R_i is the Born radius for atom i, and δ_{ij} is the Kronecker delta. This equation and similar forms that allow the original Born approach to be extended to multi-particle systems are referred to as the generalized Born (GB) equations. One such approach was proposed by Still and coworkers [69] for calculating solvation energies of organic molecules; a quantum chemical based approach was developed in the lab of Truhlar [70,71]. Still's method is based on an empirically determined functional form to calculate the polarization free energy.

The proposed function was parameterized to account for both electrostatic damping and solvation. The success of the method in calculating solvation energies of small organic molecules prompted several workers to adapt it to calculating electrostatic effects in biomolecules. The further development of this theory is summarized in several reviews and research articles [72,73], and a number of alternative models are now available: HCT [71], ACE [74], AGBNP [75,76], GBMV [77,78], GBSW [79] and ALPB [80–82].

Microscopic methods

The advantage of microscopic theory is that, in principle, no empirical parameters are needed, so that the underlying physics can be revealed. A second major advantage is that physical quantities defined at the macroscopic level, e.g., the permittivity, do not appear in microscopic formulations since the relative permittivity in a fully explicit, atomistic description is one. The major disadvantage of microscopic approaches is that they are computationally intensive, thus simplifications have to be made that can compromise the theoretical content of the method.

An important early approach in this direction was made by Warshel [83,84] who expressed the protein-solvent system in terms of charges and dipoles in the protein and point dipoles on a three dimensional grid for the solvent. Warshel's approach is based on the dielectric theory of polar solvation developed by Lorentz, Debye, Sack, and Onsager (LDSO) (see for example Ref. [85]), which, however, maintained the microscopic treatment of the entire system. Unfortunately even Warshel's approximations were still too compute-intensive so that further simplifications had to be introduced leading to a semi-microscopic approach that finally forced the reintroduction of a permittivity like quantity in the formulation.

Nevertheless, Warshel recognized that the particular form or value of the permittivity depended on the physics of the system and should not be treated as an arbitrary parameter [49,86].

The most fundamental approach for describing electrostatic, as well as all other physical interactions, are quantum mechanical (QM) methods which solve the Schrödinger equation (SE) at some level of approximation. For macromolecular systems like proteins, solving the SE for the entire system is neither possible nor desirable. The required computing power is not available, but more fundamentally, at separations where the overlap repulsion has become vanishingly small only electrostatic interactions are non-negligible and therefore must be included in the calculation. Because of these issues, most methods follow a suggestion made by Warshel and Levitt [87] to divide the system into regions where only the region of detailed interest is described by QM and the more distal parts of the system are described classically. Several such approaches are described below.

Quantum mechanics/molecular mechanics (QM/MM) based methods

A computational methodology for protein pK_a predictions, based on *ab-initio* quantum mechanical treatment of part of the protein and linear Poisson-Boltzmann equation treatment of the bulk solvent, has recently been developed by Jensen and coworkers [88]. This method was applied to predict and interpret the pK_a values of the five carboxyl residues (Asp7, Glu10, Glu19, Asp27, and Glu43) in the serine protease inhibitor turkey ovomucoid third domain and it was found to give quite promising results. Another approach described the development and application of a computational method for the prediction and rationalization of pK_a values of ionizable residues in proteins, based on *ab-initio* QM and the effective fragment potential (EFPs) method [89]. In this approach the quantum region is surrounded by fragments for which the (static) potentials have been pre-determined using *ab-initio* QM. An attractive feature of this approach is that it requires no empirical parameters[89]. It was shown that the hydrogen bonds, rather than long-range charge-charge interactions primarily determined the pK_a values. Cui and coworkers also applied QM/MM potential function in microscopic pK_a simulations[90], developing the QM/MM-GSBP[91] (Generalized Solvent Boundary Potential) based thermodynamic integration (TI) approach for pK_a predictions. The system set-up is identical to a recently published study [92] of V66E and V66D mutants, which has a 22 Å fully flexible inner GSBP region; several simulations were also been carried out with the simpler stochastic boundary condition with a large (34 Å) water sphere. To encourage structural response in the environment, the interaction between the QM titratable group and the MM environment is scaled by a constant α (>1) in the overcharging windows. Two schemes were explored: *(a)* random walks between each TI window with a specific λ value and several overcharging windows with the same λ but different α values were realized with a Landau-Wang scheme; and *(b)* random walks were realized between all TI windows and the overcharging windows; only the overcharging windows with λ=1 were included. It is clear that, while these methods show great promise, at the present stage of development further effort will be required before they can be used routinely on large sets of cases.

Molecular Dynamics (MD) based methods

In parallel to QM/MM approaches methods utilizing MD simulation have recently been proposed at various levels of approximation. These are combined with free energy perturbations (FEP) to calculate the change in free energy accompanying protonation or deprotonation. An interesting new approach carries out the simulations at constant pH allowing a first principle description of acid-base equilibria in proteins. Computational limitations require that in most applications some level of approximation is still required, which usually is achieved by using a continuum solvent approximation.

Alternative backbone conformations can be sampled within standard molecular dynamics protocols[93,94]. These approaches calculate the pK_a as a thermodynamic average from conformations in the trajectory or from an average structure. Another approach, combining both MD and the Generalized Born (GB) model, for predicting pK_as was recently reported [95]. This implementation of the Molecular-Mechanics Generalized-Born Surface-Accessibility (MM-GBSA) approach was tested on a panel of nine proteins, including 69 individual comparisons with experiment. An issue with these calculations is that values of $\varepsilon>1$ were used within the context of all atom microscopic simulations where the permittivity should be unity (Use of $\varepsilon>1$ within the context of a microscopic calculation is physically problematic). It was shown that the inclusion of non-electrostatic terms that are part of the MM-GBSA free energy expression, improved prediction accuracy. A similar observation was previously made [64] by the authors of the MCCE method concerning the inclusion of van der Waals energy into pK_a calculations. Another approach to conformational averaging is adopting a linear response approximation using conformations from both the ionized and neutral forms of the residue of interest. This approach was pioneered by Warshel within the context of the PDLD model [83] and has been recently extended to PB-based models [96,97]. Recently Washel proposed a so called overcharging approach to favor the conformational changes occurring in the MD simulations, by overcharging the titratable group of interest [98].

A method for pK_as predictions [99] was recently reported using continuous constant pH molecular dynamics (CPHMD) simulations [100,101], which employs λ dynamics for simultaneously propagating conformational and protonation states (for a review see [102]). The method calculates solvation effects using the GB model, accounts for the ion screening through approximate Debye-Hückel function and applies a replica-exchange protocol for enhanced sampling in both conformational and protonation space. By allowing the microscopic coupling between protonation equilibria and conformational dynamics, the CPHMD method offers pK_a predictions at a first-principles level, thereby eliminating the need for the effective protein dielectric constant and high-resolution structure as typically required by macroscopic approaches. Another strength of the method is that it can be applied to study pH-dependent conformational phenomena [99,102]. The CPHMD method was benchmarked on 10 proteins, targeting anomalously large pK_a shifts for the carboxylate and histidine side chains. pKa of buried ionizable groups were somewhat less well reproduced than surface groups[99]. Since the July 2009 Telluride meeting, Shen and coworkers have extended the CPHMD method to explicit-solvent simulations using a hybrid scheme in which protonation states are propagated using the GB model but conformational dynamics is driven in explicit solvent [103]. This modified method may yield an improved accuracy for the description of protein conformational dynamics while maintaining the efficiency for sampling protonation states.

An alternative constant pH approach has been developed using discrete protonation states and GB electrostatics [104]. In this method, J. Mongan *et al.* use GB-solvated MD, with periodic Monte Carlo sampling of discrete protonation states using the same GB electrostatics, to account for the important pairing of conformational dynamics and protonation state. At each MC step, a titratable residue and a new protonation state are chosen at random, with the total transition energy being used as the Metropolis criterion for the decision of protonation state. More recently, in an attempt to overcome the commonly reported convergence issues associated with constant pH MD methods, this approach has been coupled with accelerated MD.[105] Using this coupled method (CpHaMD) [106], improvement has been observed in the pK_a predictions of titratable residues of the extensively studied Hen Egg White Lysozyme (HEWL) system, relative to the earlier approach (above).

Baptista and co-workers have proposed two different constant-pH MD methods [107,108] that explore the complementarity of MM/MD methods (which sample conformations at a fixed protonation state) and PB models (which sample protonation states at fixed conformation). The first method, termed *implicit titration* [107], uses fractional protonation states periodically updated from PB calculations performed along the MD simulation. The method is based on a potential of mean force ensuring sampling from the proper semi-grand canonical ensemble, together with a mean field approximation. The second method, termed *stochastic titration* [108], uses discrete (nonfractional) protonation states which are similarly obtained from periodic PB and MC calculations. This method adopts a coupling between the MM/MD and PB/MC algorithms that generates a Markov chain sampling from the semi-grand canonical ensemble, allowing also for the use of explicit solvent in the MM/MD segments by means of an approximation; the treatment of protonatable groups with hydrogen isomerism[109] and of redox groups (by specifying the solution reduction potential) [110] was later included. The stochastic titration method succesfully reproduced the helix-coil transition of polylysine[111] and predicted the acidic pK_a values of hen egg white lysozyme in reasonable agreement with experiment[109].

Continuum methods from the microscopic description

Unlike macroscopic methods where the applicability to microscopic systems has to be assumed, continuum solvent models can be rigorously derived from the microscopic description (for a review, see Ref. [85]). Because the method is derived from microscopic electrostatics an internal dielectric constant does not appear. Instead, statistical averaging of the electrostatic equations defines a “virtual” fluid that penetrates all of space, and is described by a sigmoidal, distance dependent screening function that modulates both the electrostatic interactions and the self-energy. It provides an alternative approach for calculating pK_a that was first developed by Mehler [85]. In this approach, a variational method is used to assign the titration charge to the atoms of the titrating moiety in an optimal and self-consistent way. In a later modification, a quantitative description of the hydrophobicity of the local environment was introduced that provides a mechanism to empirically modulate the electrostatic equations based on the properties of the local environment and the degree of solvent accessibility (the method contains 5 empirical parameters). A similar approach has been reported[112] that uses the electrostatic equations derived from LDS theory, but these authors introduced empirically determined screening functions based on the region in the protein where the ionizable group is located.

Empirical methods

In contrast to the methods described above that are based on the macroscopic or microscopic electrostatic equations, the methods described here are based on an empirical functional form with parameters optimized on the basis of a large data base of measured pKa values. For example, a study that utilized a genetic algorithm to design an empirical equation that took into account the long-range charge-charge interactions and the interactions of the given carboxylic acid group with its local environment in the protein [113]. Another approach was taken by Spassov and co-workers [114], where a three terms empirical function describing charge-charge interactions was optimized over experimentally determined titration curves. Another method [115] defines an empirical equation that predicts the pK_as based on the electrostatic potential, hydrogen bonds, and accessible surface area.

A very fast and empirical method (PROPKA) was recently developed by Jensen and coworkers [116,117]. It uses the 3D structure of the protein to estimate the desolvation effects and intra-protein interactions by positions and chemical nature of the groups proximate to the pK_a sites. PROPKA was tested on 233 carboxyl, 12 cysteine, 45 histidine, and 24 lysine

pK_a values in various proteins resulted in a root-mean-square deviation less than one pH unit. PROPKA has become the most-widely used empirical program for pK_as predictions.

Recently, a new method was developed by Milletti for protein pK_a calculations, MoKaBio [118], which is based on a statistical method trained on experimental pK_a values of 434 unique residues. Each residue in the training set is described by a fingerprint that encodes the chemical environment within a sphere with a radius of 6 A from the site of ionization. This fingerprint contains information on the physical chemical properties of the neighboring atoms (charge, hydrophobicity, etc.) and their distance from the site of ionization. The prediction requires the following steps: *(a)* generation of a fingerprint for each ionizable site of a protein; *(b)* calculation of a similarity index (SI) between each fingerprint of the protein and all the fingerprints in the training set; *(c)* pK_a prediction by using experimental pK_a values of the top ten most similar ionizable sites in the training set weighted according to the SI. Leave-one-out cross-validation of this method on the training set of 434 pK_a values was carried out. In the development phase of this method it was observed that it was difficult to predict a pK_a shifts originating from long-range interactions. This motivated the authors of MoKaBio to choose a fingerprint similarity approach rather than other machine learning approaches such as Partial Least Square, which are based on the calculation of the contribution of individual groups to the pK_a shift of a residue.

INSIGHTS AND DIFFICULTIES ENCOUNTERED BY pK_a-COOPERATIVE PARTICIPANTS

The experimental dataset

The set of experimental pK_a values used for the blind prediction were obtained from crystallographic structure determinations of WT and mutants conducted by the Garcia-Moreno group. pK_a values were determined by the Garcia-Moreno group [3,5,6,119] for mutant proteins by performing equilibrium denaturation measurements at different pH and/or relevant NMR experiments. Mutants were designed to position a single ionized group in the core of SNAse to measure the effect of desolvating the ionizable group and plausible compensation from newly formed favorable interactions. This yielded highly perturbed pK_a values for a large number of residues at different positions in the sequence [2–6], which provided a unique dataset for the blind predictions. At the time of the blind prediction exercise, 90 of the mutant pK_a values had not been released and could therefore be used for a true blind prediction exercise (pK_a values were known only to the Garcia-Moreno and the Nielsen lab at the time of submission).

It is important to stress that only a single pK_a value (that of the inserted residue) was available for each mutant protein. Furthermore, for 77 of the 90 mutant proteins only modeled structures (provided by Emil Alexov) were available. In the blind prediction, each group was free to construct their own models of the mutant proteins, and the predictions submitted thus presented an exercise in both modeling and pK_a prediction. Additionally, the experimental data set is exceptional in that it contains a very large fraction of highly shifted pK_a values (average shift from the solution pK_a value is for Asp and Glu are 2.8 and 2.3 units, respectively). Finally, it should be mentioned that upon learning of their performance on the full set of pK_a values, the participants in the Telluride meeting decided to receive experimental information for only 1/3 of the full set of pK_a values. The remaining 2/3s of the pK_a values have been withheld for additional blind predictions until May 2010, and have led to improvement in the performance of some methods.

Calculations utilizing rigid heavy atom positions

The Baker/Nielsen group made predictions utilizing two protocols: PDB2PKA and WHAT IF. It was found that PDB2PKA performed particularly poorly on lysines, presumably because there was very little data on these residues in the calibration and training set. In contrast, WHAT IF yielded high RMSD for histidines in WT SNAse. Other than these observations no general trend was found in the results. However, the investigators concluded that use of a different dielectric constant would work well in improving the accuracy of some sites, while for others it appeared that one would need to explicitly sample different conformations to improve accuracy. The latter point is particularly important for the cases where only a modeled protein structure was available for the prediction, since success in the blind predictions depends crucially on calculating correctly the highly structure-sensitive desolvation energy.

The Warwicker group used a protein dielectric constant of 10 for generating predictions with the FD/DH method. The motivation of using a high dielectric constant of protein comes from the observation, that even where crystal structures are available, they may well represent non-ionised forms of the charge mutant, which upon ionization may undergo structural change. Such structural changes can be mimicked with high relative dielectrics in the range 10–12, rather than the 2–4, that are commonly used [2]. It was suggested that ionisation may introduce local conformational change, although clearly not unfolding in most cases, and predicting such conformational change is of interest. In the absence of reliable algorithms for predicting such conformational alteration, and bearing in mind that continuum models are aimed to give rapid estimates, then it may be reasonable to follow the published lead (ε_p=10) in a study focused to predict pKa of an introduced buried charge.

Calculations using rigid heavy atoms and a Gaussian model (ZAP)

Mike Word used OpenEye's ZAP PB solver to make pK_a-cooperative predictions. Although ZAP implements a discrete dielectric boundary model, its more usual mode and the mode applied here, is that of a continuous dielectric function derived from an atomic-centered Gaussian basis. This function interpolates the dielectric between the interior of the molecule and the solvent such that the predicted solvation of small molecules (<500Daltons) is within 0.5 kcal/mol of that derived from the discrete, molecular surface model of DelPhi using the same internal dielectric. There is a practical and a physical basis for this model. It is much more stable numerically, allowing estimation of solvation at an equivalent accuracy to the discrete model at about twice the grid spacing. Although it is tempting to see this model as an interpolation between the DelPhi molecular surface model and a zero-probe "van der Waals" surface model, it is actually trained to reproduce the former, i.e. to exclude water from internal spaces. However an interpolation of a kind is seen when the model is applied to larger molecules, such as proteins. As observed by Nicholls and Grant [35], calculated quantities such as binding energies, or site-site interactions are commensurate with a discrete internal dielectric, but roughly twice as large. This can be rationalized by the concept of a "wetter" protein surface than the discrete model provides and likely accounts for the correspondence between the ZAP approach and methods using a higher internal dielectric. However, there is a physical difference between the two approaches in that the underlying molecular dielectric in ZAP is still set to that from electronic polarization (ε_p=2). The higher effective dielectric occurs because the Gaussian-based function allows water more ingress to the protein, essentially sampling solvated states that might occur from small atomic displacements. In this way, the ZAP model is accounting for more than electronic polarization via the shape of the dielectric function and not from raising the intrinsic, internal dielectric. Not surprisingly, such an approach resulted in very good predictions, similar to predictions made with standard molecular surface representation and using dielectric constant of 10 for the protein.

Calculations using ensemble of backbone structures

The Alexov group applied two approaches to generate the predictions. They both were inspired by the understanding that ionization of a buried, non-paired group could induce significant conformational change. Their motivation stems from the same observation as made by Warwicker that the X-ray structures of the mutants (if available) are most probably obtained at conditions where the group of interest is not ionized (depending on the pH of the crystallography experiment). The representative structure (or ensemble of structures) with the group of interest were generated either with MD simulations or *ab-initio*. The most difficult to predict with MD generated structures were found to be Lys residues with side chain pointing directly into the hydrophobic core of the protein. The MD simulations, even up to 2ns simulation time, were not successful in generating conformational change leading to at least partial exposure of the ionized Lys side chain. On another hand, the *ab-initio* approach failed for cases where the plausible structural changes were not localized within a particular structural segment.

Explicit modeling of conformational changes through MD simulations

For the purpose of the blind predictions, Williams and co-workers utilized the constant pH MD (CpHMD) method of J. Mongan *et al.*[104] For many of the predictions, the calculated and experimentally determined pK_a results were comparable, with good representations of titration curves. However, some cases were in greater error, and the blind study highlighted some areas of the method which could be improved.

The calculation of protein pK_a values as part of a blind study was found to be more challenging. For systems where the experimental pK_a values are available, it is considerably easier to perform CpHMD simulations, since simulation length (and hence, convergence), and other method parameters can be judged, based on the known values. Williams and coworkers found convergence, an issue that was previously highlighted in constant pH MD methods, made the accurate blind pK_a prediction difficult for some residues in this study. For some of the calculations, the convergence of the pK_a value was incorrectly indicated, or was shown to be variable on performing multiple simulations. For some residues, especially those buried within the protein, strong interactions between neighboring residues persist for much of the simulation time, resulting in a low number of transitions between protonation states, and as a consequence, cause slow convergence. Therefore, simulations must generate long trajectories, and start from multiple random seeds in an attempt to help ensure that the pK_a obtained is reproducible and well-converged. However, this process was proven to be computationally expensive to carry out in a rigorous manner, especially for the numerous systems given as part of the blind study.

Since the July 2009 Telluride meeting, Williams *et al.* have adapted the CpHMD method in an effort to improve conformational sampling and thus convergence of pK_a values over simulations [106]. The CpHMD method has been coupled with the adapted Accelerated Molecular Dynamics (aMD) enhanced sampling method of de Oliveira *et al.* (described in reference [105]). This combined method (CpHaMD) employs aMD between the MC steps in replacement of conventional MD in the original CpHMD method. The use of CpHaMD has reported improvements in the pK_a predictions of the well-known problematic residues of the commonly used HEWL benchmark system, and will be further tested using the systems provided for the blind study. In addition to an increase in conformational sampling, part of the success of the method is based on the solvent model used, so any improvements made in this area would also increase the accuracy of the CpHMD method.

Shen and co-workers identified several areas of improvement. In CPHMD simulations with the GBSW implicit-solvent model [79], underestimation of effective Born radii is the main

reason for inaccuracies in the calculation of desolvation and interaction energies. The effective Born radii for buried atoms are too small because the overlapping region between van der Waals spheres that is inaccessible to water is not accounted for in the volume integration used to calculate the effective Born radii[99]. As a result, the solvation energies for buried atoms are overestimated, while the Coulomb interactions between buried sites are dampened too much. For a buried ionizable side chain, the low dielectric environment favors the neutral state while attractive electrostatics interactions with nearby groups stabilize the charged state. Underestimation of effective GB radii leads to smaller magnitude in pK_a shifts due to desolvation and due to attractive electrostatic interactions. However, because of the opposite signs, these two errors cancel each other, resulting in smaller errors in the predicted pK_a's for most interior groups, although it is not possible to predict this cancellation *a priori*

Baptista and co-workers used their stochastic titration method to run constant-pH MD simulations of just one of the mutants in the dataset, given that the method is computationally quite demanding. However, because of parameter issues for Arg and Lys residues, the runs had to be discontinued. This was the first time that Arg residues were considered as titrable in this method, illustrating how an unusual dataset can help identifying methodological issues.

Cui and co-workers reported encouraging findings for V66D, but also observed a number of limitations for their computational protocol for other cases. Analysis of the results indicated that the problem largely comes from the fact that in the exchange between λ-windows biased configurations are sampled in the low-λ windows. For example, the side chain of Asp66 becomes trapped in the solvent-exposed rotameric state even in the low-λ windows after exchanging with the high-λ (and overcharging) windows; this significantly underestimates the free energy derivatives in the low-λ windows, which leads to underestimated pK_a values. Therefore, it appears that the most serious challenges for sampling are for the intermediate λ windows. In this regard, the new GE-overcharging scheme is expected to be effective, especially, as discussed above, with its integration with ITS.

Continuum methods from the microscopic description

The Mehler group participation in the pK_a-cooperative resulted in a number of interesting cases, e.g., the coordinate file for I72E contains two coordinate sets (A and B) 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 weakly hydrophobic microenvironment while the B coordinate set defines a strongly hydrophobic local environment. This results in the pK_a value from the A coordinates to shift upwards, but not enough, while the B coordinates shift the pK_a up too much. The relatively large change in local hydrophobicity is due to the difference in solvent exposed surface area. Although this difference is not large the effect on the local hydrophobicity is large because of the very strong hydrophilic character of water. Therefore a relatively small change in solvent exposed surface area has a concomitantly large effect on the local environment leading to large changes in pK_a values. It would be of interest to carry out MD simulations on these two systems to determine if both structures converge to the same final pKa value.

Empirical models

Milletti used the MoKaBio program[118], which calculates pK_a values by using the pK_a of ionizable groups that have an environment similar to that of the residue of interest and has found that the predictions of pK_a shifts caused by an environment not encoded in the training set are challenging. It was demonstrated that MoKaBio predictions were very successful for cases resulting in high similarity index, but because most of the mutants are introduced in hydrophobic local environments, many of them could not find high enough

similarity in the training set to make successful predictions. Moreover similarity is probably not the only determinant effecting pKa prediction.

The Jensen lab used PROPKA on the Telluride data set and found that their results were of the same quality as other groups. Similar to many other groups they found most of the difficulties to be due to the significant structural rearrangement that can be expected by embedding a charge in more or less hydrophobic local environment buried in the protein, e.g. the mutants V39E and F34E. Another problem was related to predicting a reasonable averaged structure for the mutants where an x-ray structure was not available. Since PROPKA in its most common guise is an average-structure approach, it relies on being able to include structural reorganization through its parameterized effective potentials. As expected PROPKA was found to have problems for predicted geometries and was especially problematic for mutations where, e.g., the size of the mutant residue is significantly different from the WT residue, e.g., G20K and A90E. These two types of mutations may also destabilize the protein and make it more prone to partial unfolding, water penetration, and large structural changes to accommodate the new residue in predominantly its ionized form. Thus, the data set provided a good indication of how well the implicit structural reorganization works. Since PROPKA has been parameterized to pK_a values where the desolvation and electrostatic contributions are more or less in balance which is not the case in the hydrophobic local environments. Blind predicted pK_a showed that the desolvation model had been over-simplified.

FUTURE DIRECTIONS AND IMPROVEMENTS

PB methods

A major problem emerging from the Telluride meeting is the way the models address the molecular reorganization/response to ionization/deionization of the titratable residue. Most of the PB methods utilize either a rigid protein structure or allow for side-chain and hydrogen flexibility only. In this way, the corresponding model addresses the reorganization in a particularly crude way, generally representing the protein as a uniform dielectric medium, and the best results were obtained using ε_p=8–10, although some large shifts are poorly reproduced. However, the response of a protein to a charge modification in its interior is certainly inhomogeneous. Both structurally and dielectrically regions respond differentially as was demonstrated in the case of the reaction center protein[66]. The discussion led by Nathan Baker pointed out another, frequently overlooked problem, namely that there are many sets of parameters representing the radii and partial charges and the results may depend on the choice of force field parameters (see for example [120]). Another issue is the representation of the dielectric boundary between the protein and water phase, being either treated as a sharp or smooth boundary. Using non-discontinuous boundary allows the water high dielectric to permeate to some extent the protein interior, and thus to effectively reduce the desolvation cost. Such an approach is related in terms of the resulting dielectric map to the reduced probe radius (zero probe radius) proposed by Zhou to determine the molecular surface (see for example [120–122]).

MD-based methods and method utilizing alternative backbone structures

The choice of the dielectric constant that best substitutes for conformational changes should essentially vanish when all conformational reorganization is explicitly taken into account. Two approaches have emerged: *(a)* making predictions using alternative backbone structures taken either from alternative PDB files or generated *in silico* by some means, and then using these alternative structures in independent, standard PB pK_a calculations and using an averaging scheme to calculate the pKa as done by a number of researchers in the past; *(b)* generating the alternative backbone conformations using the same procedure (MD-based or

FEP) that calculates the pK_as. Obviously, the second approach is much more physically sound.

The advantage of the first approach is that it generates representative structures for charged and uncharged forms of the titratable group, and the results do not depend on the conformational path. Only the final structures are needed so that, they can be generated *ab-initio* or taken from PDB files crystallized (if any) at different conditions (pH for example). Specifically, the *ab-initio* partial structural remodeling (the hybrid-pK_a method used in Alexov's lab) has the advantage of quickly generating alternative backbone structures without being sensitive to large potential barriers separating alternative conformations. On the downside, such approaches need to make approximations to estimate the final pK_a predictions.

The explicit approaches (constant pH-MD based or FEP) are physically more sound and make fewer assumptions. The MD-based methods search conformation space with periodic sampling of protonation states using MC simulation. The main differences between these methods lie in their choice of solvent model and protocol for updating the protonation states. However, the convergence can be a problem in case of MD-based methods. Some structural relaxations may require simulations longer that several ns, or may simply be inaccessible with standard MD simulations. Enhanced sampling techniques such as replica exchange [99] and accelerated MD [106] have been employed to overcome such limitations. In addition to the sampling issue, some of the constant pH MD methods employ implicit solvation, which may limit the accuracy of pK_a predictions due to deficiencies of the solvent model in calculating electrostatic energies and sampling of conformational states. Improvements to the solvent models and/or incorporation of explicit-solvent sampling would surely increase the accuracy of these methods.

Continuum methods from the microscopic description

Unlike many pK_a programs, the MM-SCP approach of Mehler and co-workers allows the user to adjust several parameters. These include some control over the iterative process to help ensure rapid convergence. Another parameter allows damping of the electrostatic interactions below an input threshold distances. The purpose of this parameter is to partially account for cases where interatomic distances are too small. Both the threshold distance as well as the damping factor can be adjusted. In their use of the program they have found that with some experience the appropriate values of these control parameters could be estimated. Nevertheless, default values have been provided for all adjustable input parameters. A recent analysis of the method using a data base derived from 59 proteins has shown that the calculation of pKa values of histidine is the most problematic with the largest percentage of residue in error by > 1 pKa unit.

Empirical methods

The empirical methods are fast and it was found that they typically do not make large errors. This makes them ideal for quick and large-scale pK_a calculations to get an overview over up-shifted or interesting pK_a values that might be of biological importance, e.g. the two catalytic residues in lysozyme. They are unlikely to predict large pKa shift and thus to perform very well on a dataset comprised of slightly perturbed pKa's, but they will probably not pinpoint the value of "difficult" residues (large pKa shifts) in an extreme environment. In practice, this means also that they are less sensitive since they have been parameterized against predominantly near-surface residues. The most straightforward way to improve their performance in this context is to enlarge the training dataset with a diverse set of residues that include significantly shifted pK_a values. In case of PROPKA, participating in the pK_a-cooperative has already initiated such efforts and has already resulted in a better description

of the energy terms. The biggest obstacle at this point is similar to most methods discussed here, namely, how to deal with large structural reorganization (partial unfolding and water penetration). Even though it is easy to conceive approaches to include this, e.g. with MD, MC, or rotamer sampling, it would do so at the expense of its strength: computational speed and usability. The future of empirical pK_a predictors probably lies in practical use within the much larger domain of non-extreme residues and as a screening tool for more advanced methods. In case of MoKaBio, it will include more representative cases with known pK_as that will results to better similarity index (SI) and thus to more reliable predictions.

CONCLUSIONS

The pK_a-cooperative inspired 12 groups to make blind predictions for 77 experimentally determined pK_as. Due to the efforts of the Garcia-Moreno group [2–8,119,123–127], such a large benchmark of experimental pK_a values and in some cases experimentally determined X-ray structures, paved the way for broad range blind testing of a variety of methods with different physical platforms. The most striking result of this blind test was that nobody performed significantly better than the rest of the participants. Each method had successful and unsuccessful predictions, and thus indicating that all methods had problems with their underlying physics, with different problems in different methods. Much of the meeting was dedicated to discussing the reasons for this failure, with several potential reasons being pointed out as outlined above. Overall, the meeting was a great opportunity to discuss frankly the problems of the methods, which is invariably more enlightening and more productive than discussing achievements.

From the presentations and discussions of calculating pKa in proteins a steady, albeit somewhat slower than desired, improvement in accuracy can be seen. Therefore, it does not seem unreasonable to expect further progress during the next few years as methods are refined and new algorithms are proposed. If this progress is to make an impact on the Biophysics community and subsequently on the larger community of Biologists it will be necessary to become cognizant of how acid/base equilibria impact biological systems. In particular, because pKa are logarithmic quantities a shift of one pKa unit implies a ten-fold change in concentration, and given the tight control of pH in most biological systems, it is clear that a change in proton concentration implied by a shift of one pKa unit will not be tolerated by most body compartments. Thus it seems that the initial goal to strive for is to be able to predict pKa with errors < 1. Unfortunately, this means that our favorite indicator, RMSD, is of little use, since an RMSD of 0.3 does not guarantee that all pKa of a system are predicted within one pKa unit of their actual value (at least on the average). Fortunately there are many cases involving biological systems where the pKa value do not have be known to high accuracy. Instead, what is required to rationalize a biological process is to know the protonation state under a given set of experimental conditions as has been shown in a recent publication. [128]

Acknowledgments

EA and ELM thank the consortium for helpful discussions, and the thoughtful and open contributions made by the blind contributors. Some of the sections of this overview in “OVERVIEW OF METHODS FOR CALCULATING pK_as IN PROTEINS” and “INSIGHTS AND DIFFICULTIES ENCOUNTERED BY pK_a-COOPERATIVE PARTICIPANTS” are based on their contributions. Also any incorrect statements made in these sections (or any other) are entirely the responsibility of EA and ELM. Finally, the support of grants NIGMS R01GM093937, and NIH R03LM009748 (EA) and R01 DA015170 (ELM) is gratefully acknowledged.

References

1. Garcia-Moreno B. Adaptations of proteins to cellular and subcellular pH. J Biol. 2009; 8(11):98. [PubMed: 20017887]
2. Baran KL, Chimenti MS, Schlessman JL, Fitch CA, Herbst KJ, Garcia-Moreno BE. Electrostatic effects in a network of polar and ionizable groups in staphylococcal nuclease. J Mol Biol. 2008; 379(5):1045–1062. [PubMed: 18499123]
3. Castaneda CA, Fitch CA, Majumdar A, Khangulov V, Schlessman JL, Garcia-Moreno BE. Molecular determinants of the pKa values of Asp and Glu residues in staphylococcal nuclease. Proteins. 2009; 77(3):570–588. [PubMed: 19533744]
4. Isom DG, Cannon BR, Castaneda CA, Robinson A, Garcia-Moreno B. High tolerance for ionizable residues in the hydrophobic interior of proteins. Proc Natl Acad Sci U S A. 2008; 105(46):17784–17788. [PubMed: 19004768]
5. Takayama Y, Castaneda CA, Chimenti M, Garcia-Moreno B, Iwahara J. Direct evidence for deprotonation of a lysine side chain buried in the hydrophobic core of a protein. J Am Chem Soc. 2008; 130(21):6714–6715. [PubMed: 18454523]
6. Harms MJ, Schlessman JL, Chimenti MS, Sue GR, Damjanovic A, Garcia-Moreno B. A buried lysine that titrates with a normal pKa: role of conformational flexibility at the protein-water interface as a determinant of pKa values. Protein Sci. 2008; 17(5):833–845. [PubMed: 18369193]
7. Isom DG, Castañeda CA, Cannon BR, Velu PD, García-Moreno EB. Charges in the hydrophobic interior of proteins. Proc Natl Acad Sci USA. 2010; 107:16096–16100. [PubMed: 20798341]
8. Isom DG, Castañeda CA, Cannon BR, García-Moreno EB. Large shifts in pKa values of lysine residues buried inside a protein. Proc Natl Acad Sci USA. 2011 (in press).
9. Harms MJ, Castaneda CA, Schlessman JL, Sue GR, Isom DG, Cannon BR, Garcia-Moreno EB. The pK(a) values of acidic and basic residues buried at the same internal location in a protein are governed by different factors. J Mol Biol. 2009; 389(1):34–47. [PubMed: 19324049]
10. Karp DA, Stahley MR, Garcia-Moreno B. Conformational consequences of ionization of Lys, Asp, and Glu buried at position 66 in staphylococcal nuclease. Biochemistry. 49(19):4138–4146. [PubMed: 20329780]
11. Tanford C, Kirkwood JG. Theory of Protein Titration curves I. General Equations for Impenetrable Spheres. J Am Chem Soc. 1957; 79:5333–5339.
12. Bjerrum N. Dissoziationskonstanten von mehrbasischen Säuren und ihre Anwendung zur Berechnung molekularer Dimensionen. Physik Chem Stoechiom Verwandschaftsl. 1923; 106:219–241.
13. Stanton C, Houk K. Benchmarking pKa Prediction methods for residues in Proteins. J Chem Theory and Computation. 2008; 4:951–966.
14. Andre I, Linse S, Mulder FA. Residue-specific pKa determination of lysine and arginine side chains by indirect 15N and 13C NMR spectroscopy: application to apo calmodulin. J Am Chem Soc. 2007; 129(51):15805–15813. [PubMed: 18044888]
15. Gao G, DeRose EF, Kirby TW, London RE. NMR determination of lysine pKa values in the Pol lambda lyase domain: mechanistic implications. Biochemistry. 2006; 45(6):1785–1794. [PubMed: 16460025]
16. Song J, Laskowski M Jr, Qasim MA, Markley JL. NMR determination of pKa values for Asp, Glu, His, and Lys mutants at each variable contiguous enzyme-inhibitor contact position of the turkey ovomucoid third domain. Biochemistry. 2003; 42(10):2847–2856. [PubMed: 12627950]
17. Perez-Canadillas JM, Campos-Olivas R, Lacadena J, Martinez del Pozo A, Gavilanes JG, Santoro J, Rico M, Bruix M. Characterization of pKa values and titration shifts in the cytotoxic ribonuclease alpha-sarcin by NMR. Relationship between electrostatic interactions, structure, and catalytic function. Biochemistry. 1998; 37(45):15865–15876. [PubMed: 9843392]
18. Warshel A. Calculations of Enyzmatic Reactions: Calculations of pKa, Proton Transfer Reactions, and General Acid Catalysis Reactions in Enzymes. Biochemistry. 1981; 20:3167–3177. [PubMed: 7248277]
19. Click TH, Kaminski GA. Reproducing basic pKa values for turkey ovomucoid third domain using a polarizable force field. J Phys Chem B. 2009; 113(22):7844–7850. [PubMed: 19432439]

20. Mitra R, Shyam R, Mitra I, Miteva MA, Alexov E. Calculation of the protonation states of proteins and small molecules: Implications to ligand-receptor interactions. Current computer-aided drug design. 2008; 4:169–179.
21. Tanford C, Roxby R. Interpretation of Protein Titration Curves. Application to Lysozyme Biochem. 1972; 11:2192–2198.
22. Reynolds JA, Gilbert DB, Tanford C. Empirical correlation between hydrophobic free energy and aqueous cavity surface area. Proc Natl Acad Sci USA. 1974; 71:2925. [PubMed: 16578715]
23. Matthew JB, Gurd FRN, Garcia-Moreno B, Flanagan MA, March KL, Shire SJ. pH-Dependent Processes in Proteins. CRC Criti Rev Biochem. 1985; 18(2):91–197.
24. Havranek J, Harbury P. Tanford-Kirkwood electrostatics for protein modeling. PNAS USA. 1999; 96:11145–11150. [PubMed: 10500144]
25. Nicholls A, Honig B. A rapid finite difference algorithm utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J Comp Chem. 1991; 12:435–445.
26. Rocchia W, Alexov E, Honig B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. J Phys Chem. 2001; 105(85):6507–6514.
27. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid Grid-based Construction of the Molecular Surface and the Use of Induced Surface Charges to Calculate Reaction Field Energies: Applications to the Molecular Systems and Geometrical Objects. J Comp Chem. 2002; 23:128–137. [PubMed: 11913378]
28. Holst M, Baker N, Wang M. Adaptive Multilevel Finite Element Solution of the Poisson-Boltzmann Equation I: Algorithms and Examples. J Com Chem. 2000; 21(15):1319–1342.
29. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A. 2001; 98(18):10037–10041. [PubMed: 11517324]
30. Lu B, Cheng X, Huang J, McCammon JA. An Adaptive Fast Multipole Boundary Element Method for Poisson-Boltzmann Electrostatics. J Chem Theory Comput. 2009; 5(6):1692–1699. [PubMed: 19517026]
31. Lu B, Cheng X, Huang J, McCammon JA. AFMPB: An Adaptive Fast Multipole Poisson-Boltzmann Solver for Calculating Electrostatics in Biomolecular Systems. Comput Phys Commun. 181(6):1150–1160. [PubMed: 20532187]
32. Yu Z, Holst MJ, Cheng Y, McCammon JA. Feature-preserving adaptive mesh generation for molecular shape modeling and simulation. J Mol Graph Model. 2008; 26(8):1370–1380. [PubMed: 18337134]
33. Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem. 2009; 30(10):1545–1614. [PubMed: 19444816]
34. Jo S, Vargyas M, Vasko-Szedlar J, Roux B, Im W. PBEQ-Solver for online visualization of electrostatic potential of biomolecules. Nucleic Acids Res. 2008; 36(Web Server issue):W270–275. [PubMed: 18508808]
35. Grant A, Pickup BT, Nicholls A. A Smooth Permittivity Function for Poisson-Boltzmann Solvation Methods. J Com Chem. 2001; 22:608–640.
36. Bashford, D., editor. An object-oriented programming suite for electrostatic effects in biological molecules. Berlin: Springer; 1997. p. 223
37. Zhou YC, Feig M, Wei GW. Highly accurate biomolecular electrostatics in continuum dielectric environments. J Com Chem. 2008; 29:87–97.
38. Bashford D, Karplus M. pKas of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. Biochemistry. 1990; 29:10219–10225. [PubMed: 2271649]
39. Potter M, Gilson M, McCammon J. Small molecue pKa prediction with continuum electrostatic calculations. J Am Chem Soc. 1994 (in press).

40. Nielsen J, McCammon A. On the evaluation and optimization of protein X-ray structures for pKa calculations. Protein Science. 2003; 12:313–326. [PubMed: 12538895]
41. Demchuk E, Wade R. Improving the Continuum Dielectric Approach to Calculating pKa's of Ionizable Groups in Proteins. J Phys Chem. 1996; 100:17373–17387.
42. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004; 32(Web Server issue):W665–667. [PubMed: 15215472]
43. Yang A-S, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of pKas in proteins. Proteins. 1993; 15(3):252–265. [PubMed: 7681210]
44. Gilson MK. Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins. Proteins. 1993; 15(3):266–282. [PubMed: 8456096]
45. Lim C, Bashford D, Karplus M. Absolute pKa Calculations with Continuum Dielectric Methods. J Phys Chem. 1991; 95:5610–5620.
46. Antosiewicz J, McCammon J, Gilson M. Prediction of pH dependent properties of proteins. JMol Bio. 1994; 238:415–436. [PubMed: 8176733]
47. Antosiewicz J, Briggs J, Elcock A, Gilson M, McCammon J. Computing the ionization states of proteins with a detail charge model. J Comp Chem. 1996; 17:1633–1644.
48. Antosiewicz J, McCammon JA, Gilson MK. The determinants of pKas in proteins. Biochemistry. 1996; 35(24):7819–7833. [PubMed: 8672483]
49. Teixeira VH, Cunha CA, Machuqueiro M, Oliveira AS, Victor BL, Soares CM, Baptista AM. On the use of different dielectric constants for computing individual and pairwise terms in poisson-boltzmann studies of protein ionization equilibrium. J Phys Chem B. 2005; 109(30):14691–14706. [PubMed: 16852854]
50. Karshikoff A. A simple algorithm for the calculation of multiple site titration curves. Protein Eng. 1995; 8(3):243–248. [PubMed: 7479686]
51. Baptista A, Soares C. Some theoretical and computational aspects of the inclusion of proton isomerism in the protonation equilibrium of proteins. J Phys Chem B. 2001; 105:293–309.
52. Warwicker J. Improved pKa calculations through flexibility based sampling of a water-dominated interaction scheme. Protein Sci. 2004; 13(10):2793–2805. [PubMed: 15388865]
53. Nielsen J, Andersen K, Honig B, Hooft R, Klebe G, Vriend G, Wade R. Improving macromolecular electrostatic calculations. Protein Eng. 1999; 12:657–662. [PubMed: 10469826]
54. Nielsen J, Vriend G. Optimizing the Hydrogen-Bond Network in Poisson-Boltzmann Equation-Based pKa Calculations. Proteins. 2001; 43:403–412. [PubMed: 11340657]
55. You T, Bashford D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. Biophys J. 1995; 69:1721–1733. [PubMed: 8580316]
56. Beroza P, Case D. Including Side Chain Flexibility in Continuum Electrostatic Calculations of Protein Titration. J Phys Chem. 1996; 100:20156–20163.
57. Kieseritzky G, Knapp EW. Optimizing pK(A) computation in proteins with pH adapted conformations. Proteins. 2007
58. Barth P, Alber T, Harbury PB. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. Proc Natl Acad Sci U S A. 2007; 104(12):4898–4903. [PubMed: 17360348]
59. Warwicker J, Watson HC. Calculation of the Electric Potential in the Active Site Cleft due to Alpha-Helix Dipoles. J Mol Biol. 1982; 157:671. [PubMed: 6288964]
60. Warwicker J. Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. J Theor Biol. 1986; 121(2): 199–210. [PubMed: 2432357]
61. Warwicker J. Simplified methods for pKa and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. Protein Sci. 1999; 8(2):418–425. [PubMed: 10048335]

62. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. J Mol Biol. 1994; 239(2):249–275. [PubMed: 8196057]
63. Cole C, Warwicker J. Side-chain conformational entropy at protein-protein interfaces. Protein Science. 2002; 11:2860–2870. [PubMed: 12441384]
64. Alexov E, Gunner M. Incorporating Protein Conformation Flexibility into the Calculation of pH-dependent Protein Properties. Biophys J. 1997; 74:2075–2093. [PubMed: 9129810]
65. Georgescu R, Alexov E, Gunner M. Combining Conformational Flexibility and Continuum Electrostatics for Calculating Residue pKa's in Proteins. Biophys J. 2002; 83:1731–1748. [PubMed: 12324397]
66. Alexov E, Gunner M. Calculated Protein and Proton Motions Coupled to Electron Transfer: Electron Transfer from QA- to QB in Bacterial Photosynthetic Reaction Centers. Biochemistry. 1999; 38:8253–8270. [PubMed: 10387071]
67. Song Y, Mao J, Gunner MR. MCCE2: Improving Protein pKa Calculations with Extensive Side Chain Rotamer Sampling. Comp Chem. 2009; 30(14):2231–2247.
68. Hoijtink G, de Boer E, van der Meer P, Weijland W. Reduction Potentials of Various Aromatic Hydrocarbons and their Univalent Anions. Rec Trav Chim. 1956; 75:487–503.
69. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. J Am Chem Soc. 1990; 112:6127–6129.
70. Cramer CJ, Truhlar DG. An SCF Solvation Model for the Hydrophobic Effect and Absolute Free Energies of Aqueous Solvation. Science. 1992; 256:213–217. [PubMed: 17744720]
71. Giesen DJ, Storer JW, Cramer CJ, Truhlar DG. General Semiempirical Quantum Mechanical Solvation Model for Nonpolar Solvation Free Energies. n-Hexadecane. J Am Chem Soc. 1995; 117:1057–1068.
72. Bashford D, Case DA. Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem. 2000; 51:129–152. [PubMed: 11031278]
73. Onufriev A, Bashford D, Case D. Modification of the Generalized Born Model Suitable for Macromolecules. J Phys Chem. 2000; 104:3712–3720.
74. Schaefer M, Bartels C, Leclerc F, Karplus M. Effective atom volumes for implicit solvent models: comparison between Voronoi volumes and minimum fluctuation volumes. J Comput Chem. 2001; 22(15):1857–1879. [PubMed: 12116417]
75. Gallicchio E, Levy RM. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. J Comput Chem. 2004; 25(4):479–499. [PubMed: 14735568]
76. Gallicchio E, Zhang LY, Levy RM. The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. J Comput Chem. 2002; 23(5):517–529. [PubMed: 11948578]
77. Feig M, Im W, Brooks CL 3rd. Implicit solvation based on generalized Born theory in different dielectric environments. J Chem Phys. 2004; 120(2):903–911. [PubMed: 15267926]
78. Lee MS, Feig M, Salsbury FR Jr, Brooks CL 3rd. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem. 2003; 24(11):1348–1356. [PubMed: 12827676]
79. Im WP, Lee MS, Brooks CL. Generalized born model with a simple smoothing function. J Comput Chem. 2003; 24(14):1691–1702. [PubMed: 12964188]
80. Sigalov G, Fenley A, Onufriev A. Analytical electrostatics for biomolecules: beyond the generalized Born approximation. J Chem Phys. 2006; 124(12):124902. [PubMed: 16599720]
81. Gordon JC, Fenley AT, Onufriev A. An analytical approach to computing biomolecular electrostatic potential. II. Validation and applications. J Chem Phys. 2008; 129(7):075102. [PubMed: 19044803]
82. Fenley AT, Gordon JC, Onufriev A. An analytical approach to computing biomolecular electrostatic potential. I. Derivation and analysis. J Chem Phys. 2008; 129(7):075101. [PubMed: 19044802]
83. Warshel A, Russell S. Calculations of Electrostatic Interactions in Biological Systems and in Solutions. Quart Rev Biophys. 1984; 17(3):283–422.

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

85. Mehler EL. The Lorentz-Debye-Sack theory and dielectric screening of electrostatic effects in proteins and nucleic acids. Theoretical and Computational Chemistry. 1996; 3:371–405.

86. Schulz C, Warshel A. What Are the Dielectric "Constants" of Proteins and How To Validate Electrostatic Models. Proteins. 2001; 44:400–417. [PubMed: 11484218]

87. Warshel A, Levitt M. Theoretical studies of enzymatic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J Mol Biol. 1976; 103:227–249. [PubMed: 985660]

88. Li H, Robertson AD, Jensen JH. The determinants of carboxyl pKa values in turkey ovomucoid third domain. Proteins. 2004; 55(3):689–704. [PubMed: 15103631]

89. Jensen JH, Li H, Robertson AD, Molina PA. Prediction and rationalization of protein pKa values using QM and QM/MM methods. J Phys Chem A. 2005; 109(30):6634–6643. [PubMed: 16834015]

90. Li H, Robertson A, Jensen J. The Determinants of Carboxyl pKa Values in Turkey Ovomucoid Third Domain. Proteins. 2003 in press.

91. Schaefer P, Riccardi D, Cui Q. Reliable treatment of electrostatics in combined QM/MM simulation of macromolecules. J Chem Phys. 2005; 123(1):014905. [PubMed: 16035867]

92. Ghosh N, Cui Q. pKa of residue 66 in Staphylococal nuclease. I. Insights from QM/MM simulations with conventional sampling. J Phys Chem B. 2008; 112(28):8387–8397. [PubMed: 18540669]

93. Zhou HX, Vijayakumar M. Modeling of protein conformational fluctuations in pKa predictions. J Mol Biol. 1997; 267(4):1002–1011. [PubMed: 9135126]

94. Vlijmen H, Schaefer M, Karplus M. Improving the accuracy of protein pKa calculations: Conformational averaging versus the average structure. Proteins. 1998; 33:145–158. [PubMed: 9779784]

95. Kuhn B, Kollman PA, Stahl M. Prediction of pKa shifts in proteins using a combination of molecular mechanical and continuum solvent calculations. J Comput Chem. 2004; 25(15):1865–1872. [PubMed: 15376253]

96. Eberini I, Baptista AM, Gianazza E, Fraternali F, Beringhelli T. Reorganization in apo-and holo-beta-lactoglobulin upon protonation of Glu89: molecular dynamics and pKa calculations. Proteins. 2004; 54(4):744–758. [PubMed: 14997570]

97. Archontis G, Simonson T. Proton binding to proteins: a free-energy component analysis using a dielectric continuum model. Biophys J. 2005; 88(6):3888–3904. [PubMed: 15821163]

98. Kato M, Warshel A. Using a charging coordinate in studies of ionization induced partial unfolding. J Phys Chem B. 2006; 110(23):11566–11570. [PubMed: 16771433]

99. Khandogin J, Brooks CL 3rd. Toward the accurate first-principles prediction of ionization equilibria in proteins. Biochemistry. 2006; 45(31):9363–9373. [PubMed: 16878971]

100. Lee MS, Salsbury FR Jr, Brooks CL 3rd. Constant-pH molecular dynamics using continuous titration coordinates. Proteins. 2004; 56(4):738–752. [PubMed: 15281127]

101. Khandogin J, Brooks CL 3rd. Constant pH molecular dynamics with proton tautomerism. Biophys J. 2005; 89(1):141–157. [PubMed: 15863480]

102. Wallace JA, Shen JK. Predicting pKa values with continuous constant pH molecular dynamics. Methods Enzymol. 2009; 466:455–475. [PubMed: 21609872]

103. Wallace JA, Shen JK. Predicting pKa values with continuous constant pH molecular dynamics. Methods Enzymol. 2009; 466:455–475. [PubMed: 21609872]

104. Mongan J, Case DA, McCammon JA. Constant pH molecular dynamics in generalized Born implicit solvent. J Comput Chem. 2004; 25(16):2038–2048. [PubMed: 15481090]

105. de Oliveira CAF, Hamelberg D, McCammon JA. J Chem Theory Comput. 2008; 4:1516–1525. [PubMed: 19461868]

106. Williams SL, de Oliveira CAF, McCammon JA. J Chem Theory Comput. 2010

107. Baptista AM, Martel PJ, Petersen SB. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. Proteins. 1997; 27(4):523–544. [PubMed: 9141133]
108. Baptista A, Teixeira V, Soares C. Constant-pH MD method based on stochastic protonation changes. J Chem Phys. 2002; 117:4184–4192.
109. Machuqueiro M, Baptista AM. Acidic range titration of HEWL using a constant-pH molecular dynamics method. Proteins. 2008; 72(1):289–298. [PubMed: 18214978]
110. Machuqueiro M, Baptista AM. Molecular dynamics at constant pH and reduction potential: application to cytochrome c(3). J Am Chem Soc. 2009; 131(35):12586–12594. [PubMed: 19685871]
111. Machuqueiro M, Baptista AM. Constant-pH molecular dynamics with ionic strength effects: protonation-conformation coupling in decalysine. J Phys Chem B. 2006; 110(6):2927–2933. [PubMed: 16471903]
112. 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]
113. Godoy-Ruiz R, Perez-Jimenez R, Garcia-Mira MM, Plaza del Pino IM, Sanchez-Ruiz JM. Empirical parametrization of pK values for carboxylic acids in proteins using a genetic algorithm. Biophys Chem. 2005; 115(2–3):263–266. [PubMed: 15752616]
114. Spassov VZ, Kashikov AD, Atanasov B. Electrostatic Interactions in Proteins. A theoretical Analysis of Lysozyme Ionization. Biochemica et Biophysica Acta. 1989; 999:1–6.
115. Krieger E, Nielsen JE, Spronk CA, Vriend G. Fast empirical pKa prediction by Ewald summation. J Mol Graph Model. 2006; 25(4):481–486. [PubMed: 16644253]
116. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins. 2005; 61(4):704–721. [PubMed: 16231289]
117. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins. 2008; 73(3):765–783. [PubMed: 18498103]
118. Milletti F, Storchi L, Cruciani G. Predicting protein pK(a) by environment similarity. Proteins. 2009; 76(2):484–495. [PubMed: 19241472]
119. Fitch CA, Karp DA, Lee KK, Stites WE, Lattman EE, Garcia-Moreno EB. Experimental pK(a) values of buried residues: analysis with continuum methods and role of water penetration. Biophys J. 2002; 82(6):3289–3304. [PubMed: 12023252]
120. Talley K, Ng K, Shroder M, Kundrotas P, Alexov E. On the electrostatic component of the binding free energy. PMC Biophysics. 2008; (1):2. [PubMed: 19351424]
121. Dong F, Vijayakumar M, Zhou HX. Comparison of calculation and experiment implicates significant electrostatic contributions to the binding stability of barnase and barstar. Biophys J. 2003; 85(1):49–60. [PubMed: 12829463]
122. Dong F, Zhou H-X. Electrostatic contribution to the binding stability of protein-protein complexes. Proteins. 2006; 65:87–102. [PubMed: 16856180]
123. 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]
124. Dwyer JJ, Gittis AG, Karp DA, Lattman EE, Spencer DS, Stites WE, Garcia-Moreno EB. High apparent dielectric constants in the interior of a protein reflect water penetration. Biophys J. 2000; 79(3):1610–1620. [PubMed: 10969021]
125. Harms MJ, Schlessman JL, Sue GR, García-Moreno EB. Arginine residues at internal positions in a protein are always charged. 2011 inder review.
126. Cannon BR, Isom DG, García-Moreno EB. pKa values of internal Asp residues in staphylococcal nuclease. 2011 (in preparation).
127. Chimenti MS, Khangulov VS, Robinson AC, Heroux A, Majumdar A, Schlessman JL, García-Moreno EB. Structural reorganization coupled to the introduction of charge in the interior of proteins: Survey of 25 internal Lys residues. 2011 (under review).

128. Zhao G, London E. An amino acid “transmembrane tendency” scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. Protein Sci. 2006; 15(8):1987–2001. [PubMed: 16877712]