PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Optimizing pK_A computation in proteins with pH adapted conformations

Gernot Kieseritzky and Ernst-Walter Knapp*

Fachbereich Biologie, Chemie, Pharmazie/Institute for Chemistry and Biochemistry, Freie Universität Berlin, Takustrasse 6, 14195 Berlin, Germany

ABSTRACT

pK_A in proteins are determined by electrostatic energy computations using a small number of optimized protein conformations derived from crystal structures. In these protein conformations hydrogen positions and geometries of salt bridges on the protein surface were determined self-consistently with the protonation pattern at three pHs (low, ambient, and high). Considering salt bridges at protein surfaces is most relevant, since they open at low and high pH. In the absence of these conformational changes, computed pK_A^{comp} of acidic (basic) groups in salt bridges underestimate (overestimate) experimental pK_A^{exp}, dramatically. The pK_A^{comp} for 15 different proteins with 185 known pK_A^{exp} yield an RMSD of 1.12, comparable with two other methods. One of these methods is fully empirical with many adjustable parameters. The other is also based on electrostatic energy computations using many non-optimized side chain conformers but employs larger dielectric constants at short distances of charge pairs that diminish their electrostatic interactions. These empirical corrections that account implicitly for additional conformational flexibility were needed to describe the energetics of salt bridges appropriately. This is not needed in the present approach. The RMSD of the present approach improves if one considers only strongly shifted pK_A^{exp} in contrast to the other methods under these conditions. Our method allows interpreting pK_A^{comp} in terms of pH dependent hydrogen bonding pattern and salt bridge geometries. A web service is provided to perform pK_A computations.

Proteins 2008; 71:1335–1348.

Key words: pK_A; continuum electrostatics; conformational flexibility; salt bridges; pH dependence.

INTRODUCTION

The protonation states of titratable residues in proteins determine the pH dependent protein stability. They tune the energetics of electron, proton, and ion transport as well as enzymatic reactions in proteins. Hence, knowledge on the pH dependent protonation states is vital to understand protein function. Different theoretical models have been established, which try to correctly predict pK_A values in proteins and help to interpret them in the light of specific structural features in the environment of titratable residues. Most of these models[1–7] employ a semi-microscopic continuum electrostatic description of the protein based on the linearized Poisson–Boltzmann equation (LPBE). Other approaches apply a Generalized–Born (GB)[8,9] or the simpler Debye–Hückel (DH) approximation[10] as well as distance-dependent dielectric screening functions in combination with an empirical quantification of the hydrophobicity of the local environment of titratable residues.[11]

In contrast to these macroscopic protein-solvent models there are also fully microscopic approaches that deploy an explicit molecular mechanics (MM) description of protein and solvent together with a thermodynamic integration (TI) technique to obtain the free energy of deprotonation.[12,13] The linear-response approximation (LRA) of TI has been combined with a semi-microscopic solvent description based on Langevin dipoles to calculate pK_A.[14,15] Furthermore, a semi-microscopic model was applied that combined an explicit description of the molecular system for conformational sampling but a continuum solvent model to compute ionization processes.[16] Finally, there are methods that rely completely on physicochemical motivated empirical functions like PROPKA.[17]

Continuum electrostatics and conformational flexibility

Classical continuum electrostatic approaches consider microscopic conformational flexibility of molecular systems only implicitly by appropriate choices of the spatially dependent dielectric constant describing medium polarizability due to the structural variations of charged and polar groups. Accordingly, a large value, $\varepsilon_{solvent} = 80$, of the dielectric constant is used

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsors: Deutsche Forschungsgemeinschaft SFB 498 (Projects A5), the Studienstiftung des Deutschen Volkes.

*Correspondence to: Ernst-Walter Knapp, Fachbereich Biologie, Chemie, Pharmazie/Institute for Chemistry and Biochemistry, Takustrasse 6, 14195 Berlin, Germany. E-mail: knapp@chemie.fu-berlin.de

Received 20 July 2007; Revised 11 September 2007; Accepted 17 September 2007
Published online 3 December 2007 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.21820

for polar solvents like water. Placing a solute molecule from vacuum in a polar solvent results in gain of free energy. Microscopically this is due to solute–solvent interactions, which lower the total free energy of the molecular system considerably because the solvent dipoles rearrange around solute atomic charges enlarging enthalpic interactions with the solute at the expense of loss in solvent entropy. In the macroscopic continuum electrostatic model the rearrangement of solvent molecules is accounted for implicitly by creating surface charges (giving raise to the so-called reaction field) at the boundary of the solute cavity separating the low dielectric medium of the solute cavity ($\varepsilon_{protein}$) from the high dielectric medium of the solvent ($\varepsilon_{solvent}$). These surface charges are complementary in sign to the solute charges resulting in a negative contribution to the total free energy. While $\varepsilon_{solvent}$ has a clear physical meaning, the appropriate value of $\varepsilon_{protein}$ is a matter of discussion and depends on the details of the continuum electrostatic model used for the protein. In general, $\varepsilon_{protein}$ should be smaller than $\varepsilon_{solvent}$, if the protein is described explicitly with atomic partial charges and flexibility. Nevertheless, the combined action of dielectric medium with explicit protein charges and flexibility may result in a protein environment, which far from the protein surface can even be more polar than water.[18] A lower limit of the value of the dielectric constant accounting for dynamic electronic polarizability at large frequencies is $\varepsilon_{optical} = 2$ which may be used for a non-polarizable force field. The more details of a molecular system are considered explicitly in terms of atomic partial charges and, particularly, also structural flexibility the smaller can be $\varepsilon_{protein}$ to simulate the remaining details left out in modeling the protein.[19,20] Hence, for equal quality of agreement with experimental quantities like pK_A and redox potentials the electrostatic computations using smaller values of the dielectric constant are superior and structural details employed in such computations are more relevant for interpretations.

Interestingly, a value of $\varepsilon_{protein} = 4$ yielded good agreement in calculations of pK_A values but caused generally larger deviations for residues close to or at the protein surface.[5] This behavior was used to advocate the usage of larger values of the dielectric constant for solvent exposed residues[5] ($\varepsilon_{surface} = 80$ vs. $\varepsilon_{protein} = 2$–$20$) or, alternatively, for residues that exhibit larger B-factors[6] ($\varepsilon_{protein} = 8$ or 16 for small or large B-factors, respectively) implying larger conformational flexibility at the protein-solvent interfaces and for atoms with larger B-factors. In another work[10] the energetics of flexible side chains were computed using a simplified DH model with uniform $\varepsilon = 80$, while buried residues were treated conventionally using the LPBE formalism with $\varepsilon_{protein} = 4$ and $\varepsilon_{solvent} = 80$. However, optimizing the agreement with experimental pK_A by variation of the dielectric constant is problematic as it obscures the true physical nature of the discrepancy by hiding shortcomings of the computational approaches.

A fully microscopic model with consistent treatment of protein charges, dipoles and structural reorganization should be able to reproduce experimental pK_A values with low values of $\varepsilon_{protein} < 4$. Warshel *et al.*[12] pioneered the application of microscopic MM models and free energy calculations for pK_A computations which, however, may suffer from insufficient sampling because of computational limitations.[13] Using a simplified solvent model describing the average solvent polarization on a discrete grid of Langevin dipoles a value of $\varepsilon_{protein} = 4$ was sufficient[15,20] and in difficult cases $\varepsilon_{protein} = 6$ was used.[20] In this semi-microscopic PDLD/S-LRA model, which was initially developed to study redox equilibria in cytochrome c,[21] electrostatic energies were scaled by $\varepsilon_{protein}$ and averaged over MM trajectories with the residue in question in its charged and neutral state while the protonation states of all other titratable residues in the protein remained fixed (LRA cycle). In the LRA cycle the free energy was estimated by evaluating start and end point of the TI range of the MM trajectory allowing the protein structure to relax for the different charge states of the considered ionizable residue. This treatment turned out to be essential when using relatively low values for $\varepsilon_{protein}$. Ignoring protein reorganization agreement with experimental pK_A values can only be achieved by using $\varepsilon_{protein}$ between 4 and 20.[20] The averaging procedure of the LRA is done for each titratable group independently and, hence, ignores to some extent correlations between protein conformations and different protonation pattern. A correct treatment would require to perform a large number of expensive MM simulations. Pairwise charge–charge interactions of strongly coupled residues are calculated in PDLD/S-LRA by a consistent microscopic approach, while for more distant pairs of charged residues Coulomb's law with a distant-dependent dielectric constant was used.[15] The latter is done to reduce the computational effort of PDLD/S-LRA.

Recently, constant-pH molecular dynamics (MD) simulations[16,22] were developed to overcome the limitation of fixed protonation states in conventional molecular mechanics. In the implementation published by Baptista *et al.*[16] a constant temperature/pressure MD simulation with explicit water is interrupted periodically by continuum electrostatic calculations with $\varepsilon_{protein} = 2$ to calculate a new low-energy protonation micro-state based on a snapshot of the current coordinates. By stringing many cycles of MD simulation combined with continuum electrostatic computations at constant pH to track the protonation states of titratable residues, it is in principle possible to obtain their protonation probabilities and pK_A values. So far, only tests with the small organic molecule succinate were presented, probably due to CPU time limitations. Lee *et al.*[22] on the other hand extend the Hamiltonian of a molecular system by additional degrees of freedom corresponding to "titration coordinates" and applied a continuum description of the solvent based on

a GB model.[23] With a bias potential the titration coordinates are coupled to pH and propagated in time together with the atomic coordinates. Here, the solute dielectric constant was set to $\varepsilon_{protein} = 1$ (i.e. vacuum) as is done in conventional MD simulations. They applied their method to proteins like lysozyme and RNaseA and obtained reasonable agreement with experiments and other theoretical methods. However, they could not demonstrate whether and how extensive conformational sampling may actually improve pK_A predictions. Quite the contrary, pK_A values generated with the constant-pH MD approach were generally less reliable compared to results from continuum electrostatic calculations. Furthermore, an ad hoc modification of vdW-radii of carboxylate atoms was necessary.

Several different approaches were developed to combine multiple conformational states with macroscopic continuum electrostatics aimed at improving the accuracy of pK_A computation. In the earliest attempt You and Bashford[2] assumed flexible side chains that adopt local conformers independent of the configuration and charge of all other residues in the protein missing important correlations between different side chains conformers and protonation states. Based on a GB and MM description, the MM-GBSA approach[9] allows in a similar fashion for conformational relaxation of the local environment of individual titratable groups changing their protonation states, while all other titratable groups are fixed in their charge neutral reference state. Beroza and Case[3] and later Georgescu *et al.*[7] accounted for these correlations by application of a Monte Carlo method to sample the relevant combinations of side chain conformers in a pH dependent matter. However, these approaches are limited by a large albeit finite number of preselected side chain conformers that may account only approximately for the self-consistency of side chain and hydrogen atom positions and protonation states at different pH values. We name these approaches "combinatorial conformation" models.

Using such a model a benchmark was established in computing a large number of pK_A values in proteins for which experimental data are available.[7] Interestingly, these results revealed a tendency of continuum electrostatic models to overestimate interactions between ion pairs at the protein surface resulting into large negative pK_A shifts of acidic residues in these salt bridges implying large conformational changes of salt bridge geometries. Their conformations, however, failed to incorporate these effects so that the authors employed an empirical correction to remove the observed discrepancies. Ionized groups buried in proteins prefer to form salt bridges because opposite charges at close distances are mutually stabilized by strong Coulomb interactions. Hence, the conformations of both groups are unlikely to vary much at ambient pH, since the loss of entropy upon forming a strong salt bridge is more than compensated by a corresponding gain in enthalpy unless one of the salt bridge partners is charge neutralized and alternative conformations of the side chains that participated in the former ion pair become accessible. The fact that crystallography rarely provides structural information referring to very acidic or basic pH values might explain why the use of crystal structures in continuum electrostatic computations introduces a strong bias towards the ionized states of salt bridge partners.

Here, we present a novel approach to combine continuum electrostatics with multiple pH adapted conformations (PACs) to accurately compute pK_A values in proteins that completely avoid empirical corrections. This is achieved by geometry optimizing all hydrogen atom positions together with side chains of ion pairs in generating global conformations that are consistent with the most probable protonation pattern at different pH.

COMPUTATIONAL METHODS

pK_A computations in proteins with a single conformation

Details of the theory are described elsewhere.[1,24] In short, the pK_A in the protein interior is computed by adding a protein induced shift ΔpK_A, which is calculated from electrostatic energies obtained from numerical solutions of the LPBE, to the unperturbed (experimental) $pK_{A,model}$ of the titratable group in solution. Electrostatic energies comprise (i) the desolvation penalty a titratable group has to pay upon entering the protein volume, (ii) the interactions of titratable with nontitratable groups as well as (iii) interactions among titratable groups in the protein, which depend on their protonation states. The first two effects give rise to the *intrinsic* pK_A ($pK_{A,intr,\mu}$), that is, the hypothetical pK_A of a titratable group μ where all other titratable groups in the protein are in their neutral reference protonation state. The latter free energy term is decomposed into individual additive contributions $\Delta\Delta W_{\mu\nu}$ arising from pair wise Coulomb interactions between the titratable residues making it possible to write the total energy of a given protonation microstate n (defined by $x_\mu^{(n)} \in \{0, 1\}$ for all individual titratable residues μ/ν and z_μ that is $+1$ for acidic and 0 for basic residues μ including the basic residue His) as[24]

$$\Delta g^{(n)} = \sum_{\mu=1}^{N} (x_\mu^{(n)} - z_\mu) RT \ln 10 (\mathrm{pH} - \mathrm{pK}_{A,intr,\mu}) + \sum_{\mu=1}^{N} (x_\mu^{(n)} - z_\mu) \sum_{\nu=1}^{N} (x_\nu^{(n)} - z_\mu) \Delta\Delta W_{\mu\nu} \quad (1)$$

for a given absolute temperature T and pH. The energy terms $\Delta\Delta W_{\mu\nu}$ are organized in the interaction matrix $\Delta\Delta W$ and contribute to the total energy, Eq. (1), only if the considered pair of residues (μ and ν) is in the

charged non-reference protonation state. For the calculation of the pK_A of a titratable group μ the full pH titration curve of that residue is derived from repeated evaluation of the thermodynamic average

$$\langle x_\mu \rangle = \frac{1}{Z}\sum_{n=1}^{2^N} x_\mu^{(n)} \exp\left(\frac{-\Delta g^{(n)}}{RT}\right), \quad \text{with } Z = \sum_{n=1}^{2^N} \exp\left(\frac{-\Delta g^{(n)}}{RT}\right) \qquad (2)$$

giving the mean protonation probability of residue μ at a given temperature T and pH. The sums in Eq. (2) were approximated by a Metropolis Monte Carlo (MC) procedure implemented by our own software Karlsberg.[25]

Multiple local side chain conformers for pK_A computations in proteins

Previous attempts to account for conformational flexibility in electrostatic pK_A computations use local side chain conformers (so called rotamers), which are used in all combinations to form different conformations for the whole protein. Since these protein conformations are constructed by combinations of uncorrelated local side chain conformers, we call them *combinatorial conformations*. In these approaches, the pH dependent relevance of the numerous possible local conformers that are obtained by combinations of rotamers is determined in a subsequent MC procedure, where the combinatorial conformations with the appropriate rotamers are chosen.

Inspired by a previous approach from You and Bashford[2] to include explicit conformations in continuum electrostatics (see supporting information for a brief discussion), Beroza and Case[3] established a method of combinatorial conformations extending Eq. (1) such that $x_\mu^{(n)}$ now describes either a protonation or conformational state of any individual flexible residue μ. In this framework ΔΔW′ contains additional elements corresponding to interactions between pairwise combinations of local side chain configurations and a MC method is used to sample protonation and conformation microstates together.

pH adapted conformations for pK_A computations in proteins

Our approach uses a different description of conformational flexibility in proteins in terms of pH adapted conformations (PACs) of the protein. These protein conformations are generated by optimizing at a given pH value simultaneously the most probable protonation microstates with the most suitable side chain conformers and H-bond pattern. In this procedure the number of protein conformations is very small, say 5–20 depending on the protein and accuracy.

Preliminary versions of this approach were already established in previous work.[24–26] Formally, it involves a simple extension of Eq. (1):

$$\Delta g^{(n,1)} = \sum_{\mu=1}^{N} (x_\mu^{(n,1)} - z_\mu) RT \ln 10 (\mathrm{pH} - \mathrm{pK}_{\mathrm{A,intr},\mu}^{(1)}) + \sum_{\mu=1}^{N} (x_\mu^{(n,1)} - z_\mu) \sum_{\nu=1}^{N} (x_\nu^{(n,1)} - z_\mu) \Delta\Delta W_{\mu\nu}^{(1)} + \Delta G_{\mathrm{conf}}^{(1)}. \qquad (3)$$

Here, "l" denotes the protein conformation and "*n*" the current protonation micro-state. Each conformation (l) is characterized by a separate set of energy look-up tables for $pK_{A,intr}^{(1)}$ and $\Delta\Delta W^{(1)}$ as well as a specific energy difference $\Delta G_{conf}^{(1)} = G_{conf}^{(1)} - G_{conf}^{(ref)}$ between conformation "l" and an arbitrary reference conformation "ref", which is taken for the charge neutral reference protonation microstate.

While the energy look-up tables $pK_{A,intr}^{(1)}$ and $\Delta\Delta W^{(1)}$ describe the energetics of protonation change within each individual conformation, $\Delta G_{conf}^{(1)}$ involves the terms describing the energetics of conformational changes. The calculation of $pK_{A,intr}^{(1)}$ values and the elements of $\Delta\Delta W^{(1)}$ is analogous to the procedure for a single conformer. To avoid arbitrary influences of bonded energies and the van der Waals energies, which can change significantly with small variations in atomic coordinates, $\Delta G_{conf}^{(1)}$ included only electrostatic energy contributions (Coulomb and solvation energy). See the theory section of the supporting information for further details.

In analogy to other approaches electrostatic energy terms were organized in multiple look-up tables to evaluate total energies of protonation pattern and conformations with an MC procedure to compute protonation probabilities for all titratable groups and, furthermore, to select the appropriate conformations for different pH values. These probabilities represent thermodynamic averages, which simultaneously consider different protonation states and protein conformations. Thus, we have included the effect of protein reorganization explicitly and follow a concept similar in spirit as the semi-microscopic PDLD/S-LRA approach mentioned earlier.[15] In contrast to PDLD/S-LRA, however, we compute pairwise electrostatic interactions in $\Delta\Delta W^{(1)}$ consistently with the same procedure (solving the LPBE) as used for calculations of $pK_{A,intr}^{(1)}$. This is an important difference, since with protein models using a single conformation, interactions in ion pairs contained in $\Delta\Delta W^{(1)}$ have a tendency to be overestimated. Also note that in contrast to the method of combinatorial conformations our approach considers correctly modifications of the dielectric boundary because of protein conformational changes.

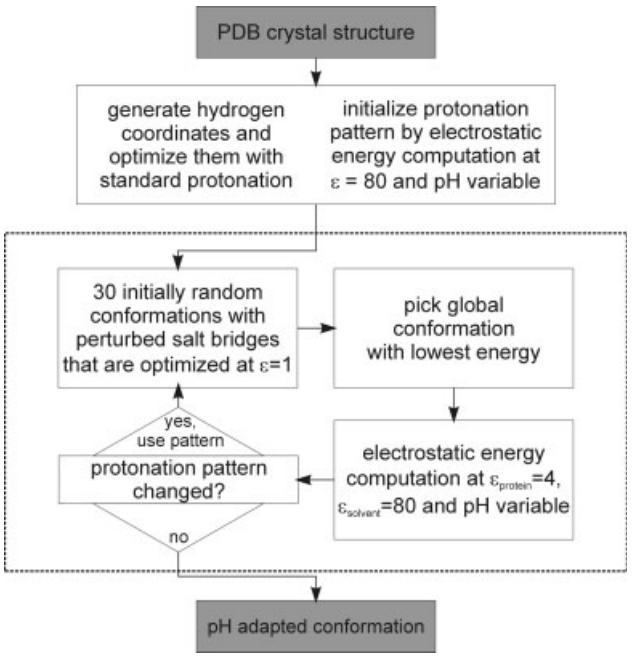

Figure 1

Workflow of the procedure to generate pH adapted protein conformations (PACs). The steps involved are the same regardless of the pH (variable pH is indicated in the steps involving electrostatic energy computations). To generate scpH7 or mc(Hflex) conformations the random search of side chain conformers is omitted and only hydrogen positions are repeatedly optimized.

Generation of PACs

The workflow is depicted in Figure 1. Initial hydrogen positions were generated using the HBUILD facility in CHARMM[27] and refined by conjugated-gradient minimization using standard protonation, that is, acids deprotonated and bases including His protonated with $\varepsilon = 1$. At this point, the protonation pattern is computed for the first time with $\varepsilon = 80$ everywhere to weaken the bias of crystal structures for pH 7. All subsequent computations of the protonation pattern are performed with an inhomogeneous dielectric with $\varepsilon_{protein} = 4.0$ and $\varepsilon_{solvent} = 80.0$. To generate appropriate pH adapted conformations (PACs) we started with a simple random search for alternative side chain conformers of residues in salt bridge geometries, which were first individually geometry optimized (with $\varepsilon = 1$ and constraining to crystal structure). These local conformers were then merged in a single global conformation and again geometry optimized (also with $\varepsilon = 1$ and constraining to crystal structure). To determine the most probable protonation state for the current protein conformation at a given pH electrostatic energy computations were performed. Then, in consecutive steps the random search of side chain conformers of residues in salt bridge geometries, geometry optimization ($\varepsilon = 1$ everywhere) with the current protonation pattern and electrostatic energy computations to determine a new protonation pattern were repeated iteratively until the protonation pattern is converged resulting in a pH adapted, self-consistent protein conformation (PAC). We applied a similar method in the past to generate relaxed protein conformations for the study of protonation and redox reactions in the bacterial reaction center.[26] However, the current protocol differs in a special treatment of salt bridges while all other residues are left invariant.

Using different seeds for random number generator typically five PACs were prepared for each of two extreme pH values (pH −8 and +20, where all titratable residues are protonated or deprotonated, respectively). For some proteins increasing the number of PACs at pH −8 from 5 to 10 improved the accuracy. These PACs were taken together with the crystal structure conformation including hydrogen positions that were determined self-consistently with the most probable protonation state at pH 7 to define a conformational ensemble as illustrated in Figure 2, which was sampled simultaneously with protonation micro-states in an MC titration.

We also did computations where we neglected conformational variations of salt bridges but still included pH dependent hydrogen flexibility (Hflex) (at pH −8, 1, 4,

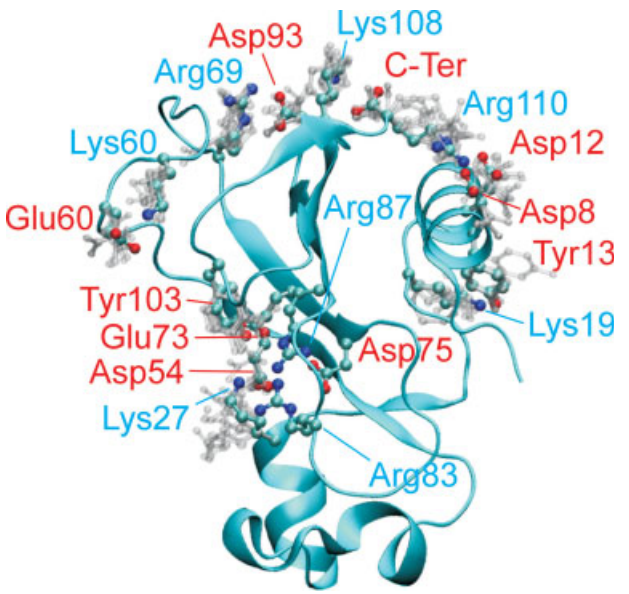

Figure 2

Ensemble of PACs used to describe conformational flexibility of barnase. Gray transparent atoms correspond to low/high pH adapted salt bridge conformations, while the crystal structure is highlighted in color. Residues that form salt bridges are labeled. The total number of conformations superimposed in this picture is 11. Three of them were energetically favored in Monte Carlo titrations (see Fig. 5). Structures were visualized using VMD47.[54]

7, 10, 13, 20) in a multiple conformation (mc) approach. We refer to these calculations by "mc(Hflex)" from now on to distinguish it from multiple conformation calculations including also optimized salt bridge (sb) geometries named "mc(Hflex,sb)". Similarly, in calculations with a single conformation (sc) called "sc_{pH7}" we prepared a single hydrogen conformation at pH 7. See details in the supporting material.

Electrostatic energy computations

The protein is described as a set of discrete partial charges in a dielectric continuum with $\varepsilon_{protein} = 4.0$ inside and $\varepsilon_{solvent} = 80.0$ outside of the protein with ionic strength 100 m*M*. The boundary between protein and solvent is defined by the solvent accessible surface area of the protein probed by a sphere of 1.4 Å radius. The atomic partial charges in the protein are all considered in detail and are identical to the charges in the CHARMM22 parameter set[28] where available. For the numerical solution of the LPBE we employed APBS 0.4.[29] For the calculation of energy look-up tables, we used our own software TAPBS, which efficiently and conveniently integrates APBS to prepare the necessary input for an MC titration as implemented in the latest 2.0 version of Karlsberg[25] directly from the three-dimensional structure of the protein and a list of titratable residues. Both Karlsberg 2.0 and TAPBS are available on our group's web site or can be used through our web interface on http://agknapp.chemie.fu-berlin.de/karlsberg that automates our PACs approach.

pK_A computations

For each PAC we computed energy lookup-tables and conformation energy differences $\Delta G^{(1)}_{conf}$, Eq. (3), as described in the supporting material. We fed the energies into Karlsberg 2.0 to compute individual titration curves of the pH dependences for all titratable groups in the given protein and evaluated pK_A^{comp} from these titration curves.

Benchmarking

For 15 different proteins 199 experimental pK_A values of titratable residues are available (see Table S4 in the supporting information for references). All protein structures used in this work were derived from structural models deposited in the Protein Data Bank (PDB).[30] For these 15 proteins 28 different crystallographic structures are available from which 22 belong to different crystals. Six additional structures are provided by multiple subunits in the same crystals. (The supporting information contains a detailed list of proteins and PDB ids.) Consequently, we could compare 391 computed pK_A values with available experimental data.

The majority of experimental pK_A (pK_A^{exp}) are very close to the corresponding model pK_A ($pK_{A,\ model}$). As pointed out by Schutz and Warshel[20] a simple correlation plot in which calculated values pK_A^{comp} are plotted as a function pK_A^{exp} will therefore show a deceivingly good agreement with experiments, because deviations from $pK_{A,\ model}$ are difficult to see on this scale (especially if values for acidic and basic residues are shown together). So we plotted instead the computed shift $\Delta pK_A^{comp} = pK_A^{comp} - pK_{A,\ model}$ as a function of the experimental shift $\Delta pK_A^{exp} = pK_A^{exp} - pK_{A,\ model}$ where all data points corresponding to residues with small ΔpK_A^{exp} will be concentrated at the origin.

A simple quality indicator is the root mean square deviation ($RMSD = \sqrt{\sum_{i=0}^{N}(pK_{A,i}^{exp} - pK_{A,i}^{comp})^2}$, the RMSD are the same if ΔpK_A is used instead of pK_A) between experimental and computed pK_A values ignoring residues where only upper or lower limits could be measured (leaving 360 computed pK_A values to compare with 185 experimental values). As explained later, we excluded the following three residues from all RMSD calculations: Asp10 (RNaseH), Asp75 (barnase) and His48 (RNaseA) (leaving 349 computed pK_A values). Furthermore, we calculated the RMSD from two subsets of data including only residues with $\Delta pK_A^{exp} \geq 1.0$ (84 computed pK_A values) and $\Delta pK_A^{exp} \geq 1.8$ (32 computed pK_A values) to discriminate between strongly and weakly shifted residues.

We also compared our approach with two other pK_A prediction algorithms, namely MCCE,[7] which is also using a continuum electrostatic description, and PROPKA[17] that is based on a physicochemical motivated empirical scoring function. In case of MCCE we took calculated pK_A values listed in the supporting material.[7] Residues present in our data set but missing in the MCCE data were not included in RMSD calculations if it came to direct comparisons. In case of PROPKA, however, we were able to use its interactive web interface[31] to compute and compare all 349 pK_A values with the results of our program.

RESULTS

pK_A computations

Table I compiles various RMSD values used in the following discussion. We achieved an RMSD of 1.12 and a maximum deviation (MAXERR) of 4.92 comparing against all experimental data [column "mc(Hflex,sb)", first row]. Ignoring experimental data with small shifts $|\Delta pK_A^{exp}| < 1$ a subset of 84 pK_A^{exp} remains for which our computations yielded a slightly lower RMSD = 1.02. Considering only strongly shifted residues obeying $|\Delta pK_A^{exp}| \geq 1.8$ our method yielded RMSD = 1.09 for the remaining subset of 32 pK_A^{exp}. Hence, the accuracy of our computations is independent of the experimental

Table I
RMSD Values and Maximum Errors (Values in Brackets) Between pK_A^{exp} and pK_A^{comp} Evaluated with Different Methods

Exp. Shift	sc_{pH7}	mc(Hflex)	mc(Hflex,sb)	NULL	PROPKA	mc(Hflex,sb)*	MCCE
all	2.66 (11.51)	1.85 (9.26)	1.12 (4.92)	1.00 (5.90)	0.89 (4.86)	1.13 (4.92)	1.04 (4.40)
$\|\Delta pK_A \geq 1\|$	3.28 (11.51)	2.19 (9.26)	1.02 (3.21)	1.79 (5.90)	1.21 (4.86)	1.06 (3.21)	1.04 (2.90)
$\|\Delta pK_A \geq 1.8\|$	3.96 (11.51)	2.89 (9.26)	1.09 (2.01)	2.38 (5.90)	1.09 (3.05)	1.11 (2.01)	1.13 (2.90)

In the first line (all) we report results for the complete set of pK_A data the second and third line refers to selections of pK_A^{exp} as indicated. All RMSD values do not include RNaseH Asp10, barnase Asp75 and RNaseA His48 for reasons discussed in text.
"sc_{pH7}" indicates computations with a single conformation and hydrogen positions consistent with the most probable protonation pattern at pH 7. "mc(Hflex)" indicates calculations with seven PACs (at pH −8, 1, 4, 7, 10, 13, 20) that vary only in hydrogen positions, while "mc(Hflex,sb)*" corresponds to computations with multiple PACs where hydrogen positions and salt bridge conformations were optimized. The pK_A^{comp} obtained with "PROPKA" were evaluated with the PROPKA web interface[12]. "NULL" corresponds to the null hypothesis where estimated pK_A are identical to $pK_{A,model}$. "MCCE" lists pK_A^{comp} taken from the supplemental material of MCCE[7] and must be compared with pK_A^{comp} of the present work listed at "mc(Hflex,sb)" which are based on a reduced set of pK_A^{comp} coinciding with the MCCE data to enable fair comparison of the methods.

pK_A shift $|\Delta pK_A^{exp}|$. This is an important finding, since for strongly shifted residues pK_A are more difficult to compute, and simple theories often fail in these cases. In this respect it is instructive to compare the results with that of the so-called "null hypothesis", where the estimated pK_A are assumed to be identical to $pK_{A,model}$. This estimate yielded for the whole data set the low RMSD = 1.0, reflecting that the majority of experimental data exhibit small pK_A shifts. However, the "null hypothesis" yielded no satisfying estimates for pK_As obeying $|\Delta pK_A^{exp}| \geq 1.0$ or $|\Delta pK_A^{exp}| \geq 1.8$ where RMSD = 1.79 and RMSD = 2.38, respectively.

How much are the results influenced by the introduction of conformational flexibility in our electrostatic approach? The top diagram in Figure 3 shows the results of calculations, which made use of only a single conformation at pH 7 (sc_{pH7}) superimposed on the results of computations that employed seven multiple pH conformations (mc(Hflex)). The sc_{pH7} conformation was generated by optimizing hydrogen coordinates in the presence of the most probable protonation pattern at pH 7. The mc(Hflex) conformations have their hydrogen positions optimized at different pH values (−8, 1, 4, 7, 10, 13, and 20) and, hence, this ensemble of seven conformations can adapt its hydrogen bonds to the prevailing protonation pattern in MC titration.

What can be seen immediately from the top diagram in Figure 3 is that the experimental pK_A shifts extend between −3 and +3, while the calculated shifts for sc_{pH7} spread between −12 and +12. The residue with the largest positive pK_A shift is Tyr53 in Omtky3 ($\Delta pK_A^{exp} = 3$) while Asp66 in lysozyme possesses the largest negative pK_A shift ($\Delta pK_A^{exp} = -3$). With sc_{pH7} ("+" symbols) the largest pK_A shifts were indeed computed for the same residues. But, they were vastly out of range with $\Delta pK_A^{comp} = +10$ and $\Delta pK_A^{comp} = -8$ for Tyr53 and Asp66, respectively. Interestingly, several acidic residues whose pK_A shift was measured to be between −1.5 and 0 were found to be much more acidic than even Asp66 in the computations. This can generally be explained by the presence of hydrogen bond donors in the proximity of these groups and, more specifically, by salt bridges that provide strong hydrogen bonds with positive charged groups. As a consequence acidic groups are exceedingly stabilized in their deprotonated states. Although this effect is real, the sc_{pH7} conformation tends to exaggerate the influence of hydrogen bonds especially on acidic groups because the hydrogen bonding pattern was generated with acidic residues in the deprotonated form. Accordingly, with sc_{pH7} the global RMSD is 2.66 and grows to 3.96 pK considering only residues with $|\Delta pK_A^{exp}| \geq 1.8$ (see Table I, column "sc_{pH7}"), which are usually found in salt bridges.

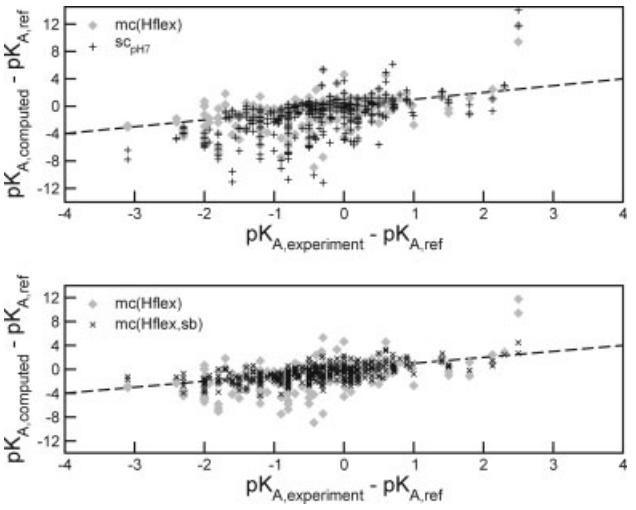

Figure 3
ΔpK_A^{comp} as a function of ΔpK_A^{exp} for different ensembles of protein conformations. Upper half: Comparison of using a single pH 7 conformation with optimized hydrogen atom positions at pH 7 (scpH[7], "+" symbols) with using conformations with pH dependent hydrogen atom positions [mc(Hflex), gray diamonds]. The RMSDs are 2.66 and 1.85 for scpH[7] and mc(Hflex), respectively. For the subset of 32 ΔpK_A^{exp} with $\Delta pK_A^{exp} \geq 1.8$ with the RMSDs are 3.96 and 2.89 for scpH[7] and mc(Hflex), respectively. Lower half: Comparison of ΔpK_A^{comp} using mc(Hflex) conformations with ΔpK_A^{comp} based on conformations that include relaxation of surface exposed salt bridges [mc(Hflex,sb)]. For mc(Hflex,sb), we obtained RMSD = 1.12. Considering only residues obeying $\Delta pK_A^{exp} \geq 1.8$ we obtained RMSD = 1.09.

In the mc(Hflex) simulations (diamonds in top diagram of Fig. 3), all hydrogen positions were modeled flexible to adapt to the protonation pattern of the protein at different pH. For example, a hydrogen bond between an acceptor such as Asp and a donor like Ser/Thr or Lys/Arg can be broken or weakened upon protonation of the acceptor. Agreement with experimental data was improved significantly with the mc(Hflex) model yielding RMSD = 1.85 for all titratable residues and RMSD = 2.89 considering only residues with $|\Delta pK_A^{exp}| \geq 1.8$, respectively [Table I, column "mc(Hflex)"]. However, many (mainly acidic) residues remained whose pK_A shifts were underestimated. This is likely due to the fact that most crystal structures are derived from crystals grown at around pH 7 where charged residues are mostly organized in ion pairs with salt bridge geometries even on the protein surface. In salt bridge geometry the ion pair constellation is particularly stable preventing a change in the protonation state even at extreme pH values.

In the lower diagram of Figure 3, we superimposed the results from calculations with PACs including optimized salt bridge geometries [mc(Hflex,sb), black crosses] with the same mc(Hflex) results shown in the upper half of Figure 3. Here, the computed pK_A shifts ranged between −4 and +4 in case of mc(Hflex,sb) and, hence, were closer to the experimental spread (±3). For lysozyme/Tyr53, ΔpK_A^{comp} improved remarkably changing from +10 to +3 (the latter is now in good agreement with experiment[32]) using mc(Hflex) and mc(Hflex,sb), respectively. With mc(Hflex,sb) conformations, RMSD = 1.12 compared to RMSD = 1.85 with mc(Hflex) for all data. This is an improvement by 0.7 pK units just by incorporating the conformational flexibility of salt bridges. Compared to the RMSD = 2.66 achieved with the single conformation sc_{pH7} we gained even 1.5 pK units with conformational flexibility. This trend is more pronounced by focusing on residues with $|\Delta pK_A^{exp}| \geq 1.8$ for which we obtained the RMSD = 1.09 using mc(Hflex,sb) in contrast to RMSD = 3.96 or RMSD = 2.89 achieved with sc_{pH7} and mc(Hflex) conformations, respectively. For titratable residues obeying $|\Delta pK_A^{exp}| \geq 1.8$, the accuracy increased by 2–3 pK units with flexible salt bridges. Furthermore, while the RMSD slightly diminished for mc(Hflex,sb) computations considering only titratable groups with $|\Delta pK_A^{exp}| \geq 1.0$, the RMSD increased by using sc_{pH7} or mc(Hflex). Most remarkably, only mc(Hflex,sb) clearly beats the "null hypothesis" for residues obeying $|\Delta pK_A^{exp}| \geq 1.0$.

Conformations

In the supporting information we describe, how we came up with a working hypothesis that salt bridges are not necessarily stable upon neutralizing the charge of one of the residues in the ion pair. This salt bridge flexibility is reflected in the PACs employed in mc(Hflex,sb) computations. Figure 4 shows conformations of three different ion pairs in the 1A2P crystal structure of barnase taken from conformations, which had the highest occupancies during MC titration. The PAC dominating between mild acidic and basic conditions, where the ion pairs form strong salt bridges, possesses coordinates of non-hydrogen atoms taken unchanged from the barnase crystal structure that was reported to be obtained at pH 7.5.[33] However, at low acidic pH the aspartates become protonated and the system prefers a conformation where the corresponding salt bridges are opened (left column of Fig. 4). Similarly, the salt bridges adopt also "open" conformations under basic conditions (right column of Fig. 4), which differ from the "open" conformations at acidic pH.

In case of the ion pair Asp93/Arg69 in barnase for example, at acidic pH Arg69 retains a single strong

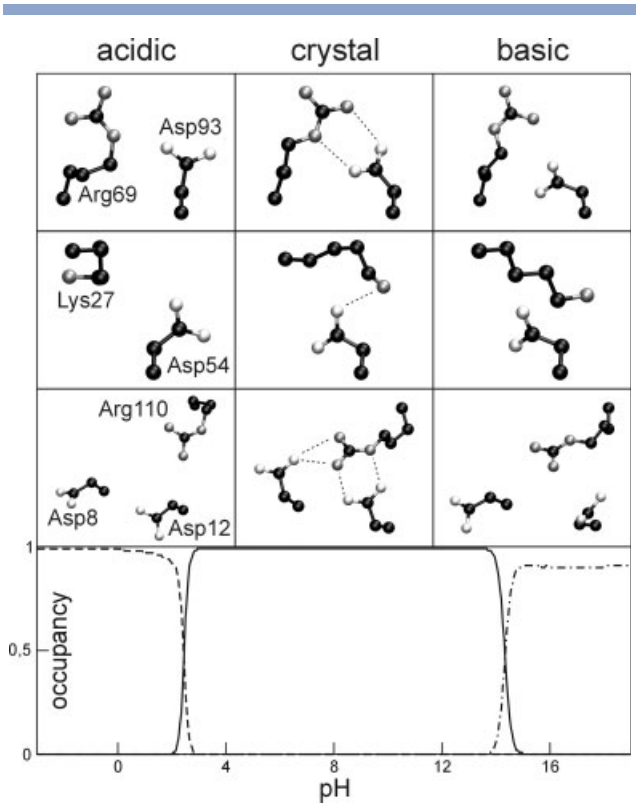

Figure 4

pH dependent conformational changes calculated for the 1A2P crystal structure of barnase. The picture shows ball-and-stick models of three different salt bridges from the same three dominant conformations that are sampled during MC titration at acidic, neutral and basic pH regions. Carbon is painted in black, nitrogen in gray, and oxygen in white. In the lower third the occupancy of the corresponding conformations, that is, the probability of the system resting in that conformation, is plotted as a function of pH. Structures were visualized using VMD.[54]

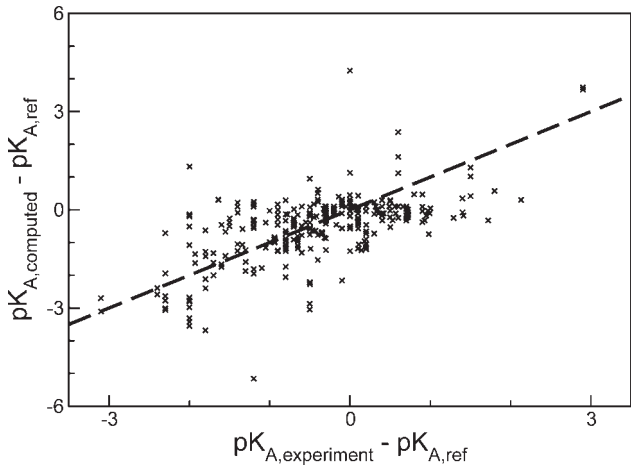

Figure 5
ΔpK_A^{comp} *evaluated by PROPKA as a function of* ΔpK_A^{exp}.

hydrogen bond (heavy atom distance < 3.0 Å) to the unprotonated oxygen in Asp93 via its NE nitrogen under acidic conditions, while at basic pH the neutral Arg69 forms only a weak hydrogen bond to Asp93 (>3.0 Å) via NH2 and two strong hydrogen bonds to the backbone carbonyl group of Pro64. Lys27 in barnase forms a salt bridge to Asp54 at ambient pH, at strongly acidic pH it donates a hydrogen to form an H bond to the backbone carbonyl group of Gln31, while at strongly basic pH it does not appear to have specific interactions with the protein but rather protrudes into the solvent.

The example of barnase is representative for the results we got for the other protein structures. However, it is important to note that the calculated conformations can not directly be considered as the "true" protein conformations at these extreme pH values but are probably similar to low temperature conformations at acidic or basic pH, respectively, since the conformations were generated by energy minimization. The important feature of these conformations is the opening or weakening of the salt bridges and in this sense they could be interpreted as pH dependent structure predictions. It is noteworthy that at low or high pH the original salt bridge conformations do not hold upon energy minimization in spite of constraints that try to maintain the coordinates of the crystal structure throughout all the energy minimization procedures. Furthermore, the average energy of micro-states sampled during MC titration is significantly smaller than the corresponding value obtained with sc_{pH7} if any of the extreme pH conformations is predominant. We found a similar effect (with varying energy differences) for all other proteins included in the test set. This supports the idea that at pH values where salt bridges "open" because of charge neutralization of one member in the ion pairs the protein relaxes into a lower energy conformation.

DISCUSSION

Comparison with other approaches

MCCE

MultiConformation Continuum Electrostatics (MCCE) is an approach developed by Georgescu *et al.*[7] based on work of Beroza and Case[3] to extend Poisson–Boltzmann continuum electrostatics to explicitly consider conformational flexibility. Starting from a given protein structure MCCE generates additional side chain conformers for polar and charged groups based on libraries that contain the most probable side chain rotamers for a given protein backbone configuration.[34] Essentially, all different conformers of a side chain are merged to define a single dielectric boundary between protein and solvent, which is applied for all different conformers. Solutions of the LPBE with partial charge distributions for different protonation and conformation states were combined to generate a united list of intrinsic pK_A ($pK^{\mu}_{A,intr}$) and a united interaction matrix $\mathbf{\Delta\Delta W}$ containing pairwise interactions between protonation states and between local side chain conformers. In addition, nonelectrostatic energy terms (Lennard-Jones potentials) are added to filter out sterically unfavorable linear combinations of local conformers during Monte Carlo pH titration.

Using the data included in the supporting material of MCCEs primary reference[7] the RMSD for all titratable residues was calculated to be 1.04. For residues with $|\Delta pK_A^{exp}| \geq 1.0$ MCCE also yields RMSD = 1.04. These RMSD are both slightly smaller than in the present PAC approach using the same set of residues [RMSD = 1.13 and 1.06, respectively, see Table I, column "mc(Hflex,sb)*"]. But, for residues obeying $|\Delta pK_A^{exp}| \geq 1.8$ MCCE yields RMSD = 1.13 while the PAC approach yields a RMSD = 1.11 for the same set of proteins.

The authors of MCCE report an RMSD of 0.82,[7] but this value is based on a subset of their data where pK_A values were specifically selected, if more than one structure of the considered protein is available to compute the pK_As. We took instead the pK_A values of all protein structures from the supplemental material[7] to calculate the RMSD (except for the pK_A of Asp10 in RNaseH, Asp75 in barnase and His48 in RNaseA for reasons discussed later).

To achieve this quality of accuracy the MCCE developers used besides conformational flexibility also spatially dependent dielectric constants. Facing the problem of over-stabilization of ion pair states in strong salt bridges, they rescaled both intrinsic pK_A values and pairwise interactions using an empirical correction function designed to weaken strong electrostatic interactions of charge pairs at short distances. This procedure corre-

sponds to a distance dependent dielectric constant, which increases at short distances of charge pairs and thus accounts effectively for additional molecular flexibility in these regions of the protein where evidently the combinatorial conformations of MCCE do not provide sufficient structural flexibility. In contrast, the PAC method considers conformational solely flexibility for ion pairs and hydrogens and no increase of the dielectric constant at short distances of charge pairs. Nevertheless, essentially the same level of accuracy is obtained as with MCCE. Without the empirical correction for the MCCE approach two different values were given for the RMSD errors, which are 1.68 (p. 1738, right column of Ref. 7) and 2.16 (p. 1739, Table 6 of Ref. 7). The first value is twice as large as the value of 0.82 pK units reported as the accuracy of MCCE including empirical corrections. The second value of RMSD error is very close to 2.29 pK given as the accuracy achieved with their value for using a single conformation (Table 2 of Ref. 7).[7] In any case, the empirical corrections of varying the dielectric constant used in the MCCE approach seem to contribute considerably to the agreement with experimental data. Hence, many of the combinatorial conformations generated by MCCE are not very useful to describe the pH dependent conformational changes in a protein appropriately. Indeed, the authors report that only 4% of the polar residues actually change their conformation significantly throughout the benchmark calculation (p. 1740, left column Ref. 7).

In summary, we found that with respect to pK_A prediction MCCE is essentially equivalent to our PAC approach. However, a closer analysis revealed that the MCCE performance in pK_A prediction is mainly based on an empirical correction, which reduces strong electrostatic interactions at short distances. These interactions are most pronounced in ion pairs. The combinatorial conformations used by MCCE are not sufficiently flexible to take care of these effects directly. This may be due to the fact that rotamer libraries as applied by MCCE to generate side chain conformations are ultimately based on crystal structures, which are obtained at pH 7 for the most part, so that these combinatorial conformations may reflect a bias towards closed conformations of ion pairs.

One essential difference between MCCE and our PAC approach is a systematic construction of a small number of pH adapted conformations that are self-consistent with the most probable protonation patterns at different pH. In these conformations only hydrogen positions and side chain conformers of ion pairs are varied to retain most of the crystal structure information. In this way, we succeed in identifying conformations that are relevant to specific pH values. Furthermore, the strategy to use a small number of pH adapted conformations instead of a large number combinatorial conformations avoids meddling with the dielectric boundary, which may be possible source of error in MCCE calculations. The combination of these arguments might explain why the PAC approach works reasonably well without using empirical corrections.

PROPKA

PROPKA[17] is a program for very fast pK_A predictions. In PROPKA the estimated pK_A shift is expressed as a sum of physicochemical motivated empirical terms. It accounts for number and geometry of hydrogen bonds to a titratable group, for desolvation effects of buried and partly buried titratable groups and distant dependent charge–charge interactions. If two titratable residues are at close distance, always the residue being more acidic is assumed to be deprotonated. The relative contributions of the individual terms to the total pK_A shift experienced by a titratable group in a protein were fitted to a representative number of experimental pK_A in proteins to give good agreement with experimental data. Using the PROPKA web interface we applied the program on our test set of proteins and compared the computed pK_A directly with ours.

PROPKA is not only fast but also outperforms both the PAC and MCCE approach for the overall RMSD, which was calculated to be 0.89 for our test set of proteins (Table I, column "PROPKA") and 0.87 for the MCCE set (not shown in Table I). However, for titratable residues obeying $|\Delta pK_A^{exp}| \geq 1.0$ PROPKA yielded a significantly larger RMSD of 1.21. The residue causing the largest error in the PROPKA prediction was His12 in RNaseA, which is affected by ion binding as discussed below. Ignoring this problematic residue we obtained for the complete set of pK_A and for the pK_A with $|\Delta pK_A^{exp}| \geq 1.0$ RMSD = 0.83 and RMSD = 1.00, respectively. Focusing only on titratable residues obeying $|\Delta pK_A^{exp}| \geq 1.8$ (which also excluded His12 in RNaseA) PROPKA achieved an RMSD of 1.09 pK. Hence, the prediction quality of PROPKA becomes significantly worse for larger $|\Delta pK_A^{exp}|$. In contrast, we found our program to perform better for larger $|\Delta pK_A^{exp}|$ with the RMSD decreasing from 1.12 to 1.02 for $|\Delta pK_A^{exp}| \geq 1.0$.

Figure 5 provides a clue to understand the characteristics of PROPKA. Many points in the correlation plot of Figure 5 lie on a horizontal line close to zero exhibiting a bias of PROPKA to predict vanishing pK_A shifts. This effect may in part be traced back to threshold values used for the individual terms that are employed to estimate pK_A shifts. Since most titratable residues investigated experimentally indeed experience nearly vanishing net pK_A shifts caused by the protein environment, PROPKA gains an advantage. Recall that applying the "null hypothesis", where one simply uses the model pK_A values, on our test set we obtained RMSD = 1.00 exhibiting a performance comparable to the other methods discussed in this study.

Analysis of individual titration sites

The aim of this work was to establish a general approach for accurate pK_A computations independent of experimental knowledge about conformational flexibility. Although a close investigation of the individual groups of all considered proteins of the test set is beyond the scope of this work, we comment results for selected sites to demonstrate how the generated PACs can be interpreted in terms of pH dependent structural variations regarding salt bridge geometries, protonation pattern and hydrogen atom positions. More examples are given in the supporting material as well as an exhaustive list of titratable residues (Table S3) whose pK_A^{comp} improved considerably upon including conformational flexibility of PAC. All pK_A^{comp} given in the following were computed with mc(Hflex,sb) if not otherwise stated.

Lysozyme Asp66

With $pK_A^{exp} = 0.9 \pm 0.5$ Asp66 possesses the lowest measured pK_A[35] of the considered data set of 22 proteins. Computations with the single sc_{pH7} conformation yielded $pK_A^{comp} = -2.4$ pK, which is more than 3 pK units lower than pK_A^{exp} due to strong hydrogen bonds with Thr69 and Tyr53. In mc(Hflex) computations, we achieved perfect agreement obtaining pK_A^{comp} of 0.9 and 1.2 with the crystal structures 2LZT and 1B0D, respectively. It turned out that Thr69 donates a pH dependent hydrogen bond to either Asp66 (pH 7) or the backbone of Gly49 (pH < 1).

Lysozyme Tyr53

Using the single sc_{pH7} conformation based on the 1B0D and 2LZT crystal structures we calculated a very basic pK_A^{comp} between 22 and 24 for Tyr53, in conflict to a moderately basic $pK_A^{exp} = 12.1$.[36] Using mc(Hflex,sb) we calculated a pK_A^{comp} of 14.1 (2LZT) and 12.3 (1B0D) in good agreement with experiment. Upon deprotonation of Tyr53 the side chains of Tyr53 and Asp66 reorient to increase the distance between this ion pair.

Xylanase Glu78 and Glu172

Glu78 and Glu172 are in the catalytic center of Xylanase. After binding the polysaccharide substrate they are located on the upper and lower side of the six-membered sugar ring, respectively. The supposed catalytic mechanism[37] involves two steps, a nucleophilic substitution of the leaving group to break the β-1,4-glycosidic bond that is followed by hydrolysis of the sugar moiety. Based on its $pK_A^{exp} = 4.6$[38] Glu78 was identified as the nucleophilic agent. For Glu78 we computed $pK_A^{comp} = 3.6$. Using the crystal structural conformation sc_{pH7} alone, we obtained $pK_A^{comp} = -0.2$ due to the fact that in the crystal structure Glu78 is involved in two short hydrogen bonds donated by Gln127 and Tyr69 with 2.7 Å heavy atom distance in both cases. Conformational flexibility of Tyr69 explains the 4 pK units discrepancy between mc(Hflex,sb) and sc_{pH7}. Glu172 plays the role of an acid/base catalyst facilitating hydrolysis of the intermediate consistent with an elevated $pK_A^{exp} = 6.7$[38] compared with the Glu model pK_A of 4.4. We computed $pK_A^{comp} = 6.8$ using the substrate-free 1XNB crystal structure in excellent agreement with these findings. From experiments with trapped reaction intermediates where Glu78 is glycosylated it has been concluded that electrostatic repulsion from Glu78 increases the Glu172 pK_A by about one pK unit.[39] We mimicked glycosylation of Glu78 by fixing it in its neutral state during Monte–Carlo titration and calculated $pK_A^{comp} = 5.7$ for Glu172 in this case compared to 6.8 when Glu78 is ionized reproducing nicely also this measured pK_A shift.

RNaseA His12

The pK_A of His12 in RNaseA is affected by phosphate/substrate binding and, hence, two pK_A^{exp} were measured: (α) $pK_A^{exp} \approx 7.0$ and (β) $pK_A^{exp} \approx 5.8$ with and without phosphate, respectively. These pK_A^{exp} are the averages of several experimental values, please see table S4 for references. It was found that phosphate binding goes along with conformational variation of the Lys41 side chain.[40,41] Our computations aimed to reproduce case β based on the 1.5 Å resolution 3RN3 crystal structure,[42] from which we removed the sulfate ion (replacing phosphate in the functional RNaseA) and alternatively based on the 1.3 Å resolution structure 7RSA,[43] which possesses neither phosphate nor sulfate group. To our surprise using 3RN3 we computed $pK_A^{comp} = 5.6$ for His12 in excellent agreement with the phosphate-free pK_A^{exp}, while we obtained a too low value of $pK_A^{comp} = 2.6$ using the 7RSA structure. The major difference between the 3RN3 and 7RSA crystal structures is the Lys41 conformer, which in case of 7RSA is closer to His12 (4.9 and 4.2 Å distances between Lys41-NZ and His12-CE1 for 3RN3 and 7RSA, respectively) explaining in part the variation in His12 pK_A^{comp}.

RNaseA His119

According to structural studies[40–42,44] His119 adopts side chain conformers with varying rotation around the $C_\alpha—C_\beta$ bond where it either forms a hydrogen bond with Asp121 (conformer A at high pH) or points into the catalytic site (conformer B at low pH). In our computations, we started from the A conformer of His119 and were able to generate a PAC at low pH with His119 similar to the B conformer (Fig. S3) and Asp121 protonated. This PAC was occupied below pH 4.0 (to 10–13% between pH 0–2 and to 5% at pH 3.5) where Asp121 was protonated. We also investigated the conformational equilibrium of His119 using solely the crystal structures and found that transition from conformer A to B goes

along with protonation of Asp121, in agreement with our PAC based computations. We computed pK_A^{comp} = 5.6 and 6.1 for His119 using 7RSA and 3RN3, respectively independent of its conformer and in good agreement with pK_A^{exp} = 5.6–6.3 (see table S4 for references).

RNaseH Glu48

For Glu48 in RNaseH we obtained three vastly different pK_A^{comp} (3.7, −0.5, 6.7) using either the crystal structure 1RNH[45], 2RN2,[46] or 1RDD,[47] where the latter involves an ion ligated to Glu48, Asp10 and Asp70, while the measured pK_A^{exp} = 4.4 is independent of Mg^{2+} binding.[48] In the 2RN2 crystal structure, Asn44 lies very close to Glu48 (2.4 Å between Glu48-OD1 and Asn44-ND2) for which it seemingly acts as hydrogen bond donor (albeit the distance is much too short). After flipping the CNO group of Asn44 in 2RN2 and minimizing the Glu48 carboxylate oxygens in their deprotonated state we obtained pK_A^{comp} = 1.9, which is closer to pK_A^{exp} = 4.4 (but which we have not used in the calculation of the RMSD). In the 1RDD crystal structure co-crystallized with Mg^{2+} Glu48 is a Mg^{2+}-ligand and the hydrogen bonds with Asn44 and Ser71 are disrupted. Here, we obtained pK_A^{comp} = 6.7 for Glu48 without Mg^{2+}, which is considerably higher than pK_A^{exp} = 4.4. After including Mg^{2+} we obtained pK_A^{comp} = −1.4 overestimating the influence of ion binding probably because we used a too simple charge model for the ligated Mg^{2+}. In a more realistic charge model beyond the scope of this study part of the twofold positive Mg^{2+} charge will be to the ligands and diminish the influence of the Mg^{2+} ion on the pK_A shift.

RNaseH Asp10

We computed for Asp10 in RNaseH pK_A^{comp} = 11–13 overestimating the experimental value by more than 5 pK (pK_A^{exp} = 6.1[48]) in absence of Mg^{2+}. On the other hand, we reproduced the biphasic titration curve of Asp10 observed in experiments,[48] where deprotonation of Asp10 is halted by deprotonation of Asp70 and Asp134. Strong interactions with Asp70 and Asp134 are probably the reason why we overestimated the pK_A^{exp} of Asp10. Here, our model does not account enough for conformational flexibility. Georgescu *et al.*[7] also reported a too high pK_A^{comp} for Asp10 using MCCE and therefore excluded it from the RMSD, which we also did for the sake of fair comparison. With Asp10 included we obtained RMSD = 1.21 as compared with 1.12 without. Interestingly, pK_A^{comp} values for Asp10 estimated by PROPKA are also larger than the experimental value but deviate only by 1 unit.[17]

Barnase Asp75

The buried residue Asp75 in barnase forms a strong bifurcated hydrogen bond with Arg83 and lies close to Arg87 suggesting a very acidic pK_A. Accordingly, we obtained a pK_A^{comp} between −4 and −2. So it came to our surprise to find pK_A^{exp} = 3.1.[49] However, in the same publication the authors report an NMR study on the pH dependent unfolding rate constant of native barnase and its Asn75 mutant where they detected an excess charge on Asp75 in the native protein at a pH as low as 0.2. This result is compatible with a pKA $\leq$ 0.2 for Asp75. Because of these experimental ambiguities we decided to exclude this residue from the RMSD. We obtained RMSD = 1.39 including Asp75 (RMSD = 1.12 without).

RNaseA His48

With pK_A^{comp} = 10.7–12.5 we significantly overestimated available experimental pK_A^{exp} = 6.2-6.4[50–52] of His48 in RNaseA. Interestingly, also MCCE[7] overestimated this pK_A significantly, although the difference is there only 1.2–1.7 units.[7] On the other hand, PROPKA underestimates this pK_A yielding pK_A^{comp} = 3.2–4.7 because of a large desolvation penalty assigned to His48. As indicated by early NMR experiments[53] and established by a recent pH dependent crystallographic study,[40,41] titration of His48 is associated with a local conformational change at Gln101, which we do not account for in our PACs approach. Clearly, this residue deserves a detailed investigation including experimental conformations, which is beyond the scope of this work. So far we removed this residue from the RMSD. We obtained RMSD = 1.25 including His48 (1.12 without).

CONCLUSION

We developed a general and robust algorithm to compute and predict pK_A values of titratable groups in proteins. The approach combines continuum electrostatic theory with multiple pH adapted conformations (PACs) generated by self-consistent geometry optimization of hydrogen positions and salt bridges at the protein surface and protonation pattern at three different pH values (low, ambient, high). In contrast to similar approaches using many uncorrelated conformers from a database of amino acid side chain rotamers the present approach employs only a few PACs, which are energetically most favorable in a specific pH range. These PACs allow interpreting specific pH dependent structural variations. The present approach demonstrates that salt bridges at the protein surface have to be open at both low and high pH to yield pK_A^{comp} of these groups in close agreement with experiment.

Considering a large set of 185 pK_A^{exp} from 15 different proteins, we have demonstrated that our PACs approach dramatically improves the accuracy of computed pK_A values compared to computations based on the usage of a single conformation derived from the crystal structure

and the protonation pattern at pH 7. The results of the present approach reach comparable accuracy with experiment as the most successful alternative approaches from which PROPKA[17] employs a fully empirical scheme to compute pK_A values with a large number of adjustable parameters while in the MCCE approach the electrostatic energies were evaluated by solving the Poisson equation with locally variable dielectric constants in the protein to diminish electrostatic interactions at short distances. We like to emphasize that we achieved this level of accuracy with a consistent use of $\varepsilon_{protein} = 4$ implying that our model captures the relevant details of the microscopic protein reorganization. At the same time our PACs approach albeit slower than a single conformation LPBE model or the fully empirical PROPKA algorithm reasonably compromises between a rigorous physical description and computational efficiency in contrast to very expensive all-atom treatments based on TI.

Future work will have to focus on titratable groups that are influenced by ion binding as they represented the greatest source of error of the present approach. However, we are pessimistic as to which extent such phenomena can be incorporated in a fully automatic pK_A prediction scheme based on continuum electrostatics, since its reliability also depends on the quality of the atomic charge models of the molecular groups involved. In particular for coordinated metal centers we would need individual and careful quantum chemical computations for atomic charges.

A web interface is available to the scientific community on: http://agknapp.chemie.fu-berlin.de/karlsberg

ACKNOWLEDGMENTS

The authors thank Drs. Martin Karplus and Nathan Baker for providing the programs CHARMM and APBS, respectively.

Note added in proof. After acception of the manuscript we learned about another empirical approach to predict pKa values in proteins (Proteins 2007;69:75–82). Applying this procedure to our test set with all 185 measured pKa we obtained the RMSD of 0.92 pK units relative to experiment. Considering only titratable groups with large pKa shifts, however, the RMSD is 1.85 as compared to 1.09 with our method (see last line of table 1 in the column labeled "mc(Hflex,sb)") demonstrating the importance of pKa computations based on physical principles.

REFERENCES

1. Bashford D, Karplus M. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
2. You TJ, Bashford D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. Biophys J 1995;69:1721–1733.
3. Beroza P, Case DA. Including side chain flexibility in continuum electrostatic calculations of protein titration. J Phys Chem 1996; 100:20156–20163.
4. Antosiewicz J, McCammon JA. The determinants of pKas in proteins. Biochemistry 1996;35:7819–7833.
5. Demchuk E, Wade RC. Improving the continuum dielectric approach to calculating pKas of ionizable groups in proteins. J Phys Chem 1996;100:17373–17387.
6. Nielsen JE, Vriend G. Optimizing the hydrogen-bond network in poisson-boltzmann equation-based pKa calculations. Proteins: Struct Funct Genetics 2001;43:403–412.
7. Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculation pKas in proteins. Biophys J 2002;83:1731–1748.
8. Luo R, Head MS, Moult J, Gilson MK. pKa shifts in small molecules and HIV protease: electrostatics and conformation. J Am Chem Soc 1998;120:6138–6146.
9. Kuhn B, Kollman PA, Stahl M. Prediction of pKa shifts in proteins using a combination of molecular mechanical and continuum solvent calculations. J Comput Chem 2004;25:1865–1872.
10. Warwicker J. Improved pKa calculations through flexibility based sampling of a water-dominated interaction scheme. Protein Sci 2004;13:2793–2805.
11. Mehler EL, Guarnieri F. A self-consistent, microenvironment modulated screened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. Biophys J 1999;75: 3–22.
12. Warshel A, Sussman F, King G. Free energy of charges in solvated proteins: microscopic calculations using a reversible charging process. Biochemistry 1986;25:8368–8372.
13. Simonson T, Carlsson J, Case DA. Proton binding to proteins: pK_a calculations with explicit and implicit solvent models. J Am Chem Soc 2004;126:4167–4180.
14. Lee FS, Chu ZT, Warshel A. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIX programs. J Comput Chem 1992;14:161–185.
15. Sham YY, Chu ZT, Warshel A. Consistent calculations of pKa's of ionizable residues in proteins: semi-microscopic and microscopic approaches. J Phys Chem B 1997;101:4458–4472.
16. Baptista AM, Teixeira VH, Soares CM. Constant-pH molecular dynamics using stochastic titration. J Chem Phys 2002;117:4184–4200.
17. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. Proteins: Struct Funct Bioinformat 2005;61:704–721.
18. Warshel A, Åqvist J, Creighton S. Enzymes work by solvation substitution rather than by desolvation. Proc Natl Acad Sci USA 1989; 86:5820–5824.
19. Warshel A, Sharma PK, Kato M, Parson WW. Modeling electrostatic effects in proteins. Biochimica et Biophysica Acta 2006;1764:1647–1676.
20. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins: Struct Funct Genetics 2001;44:400–417.
21. Langen R, Brayer GD, Berghuis AM, McLendon G, Sherman F, Warshel A. Effect of the Asn52->Ile mutation on the redox potential of yeast cytochrome c. J Mol Biol 1992;224:589–600.
22. Lee MS, Freddie R Salsbury J, Charles L Brooks I. Constant-pH molecular dynamics using continuous titration coordinates. Proteins: Struct Funct Bioinformat 2004;56:738–752.
23. Im W, Lee MS, C L Brooks I. Generalized born model with a simple smoothing function. J Comput Chem 2003;24:1691–1702.
24. Ullmann GM, Knapp EW. Electrostatic models for computing protonation and redox equilibria in proteins. Eur Biophys J 1999;28: 533–551.
25. Rabenstein B, Knapp EW. Calculated pH-dependent population and protonation of carbon-monoxy-myoglobin conformers. Biophys J 2001;80:1141–1150.

26. Rabenstein B, Ullmann GM, Knapp EW. Calculation of protonation patterns in proteins with structural relaxation and molecular ensembles—application to the photosynthetic reaction center. Eur Biophys J 1998;27:626–637.
27. Bernard R. Brooks REB, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4: 187–217.
28. MacKerell A,Jr, Bashford D, Bellott M, Dunbrack R, Jr, Evanseck J, Field M, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau F, Mattos C, Michnick S, Ngo T, Nguyen D, Prodhom B, Reiher W, III, Roux B, Schlenkrich M, Smith J, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins (Charmm22 force field). J Phys Chem B 1998;102:3586–3616.
29. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA. Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA 2001;98:10037–10041.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
31. Curtis DE, Dolinsky T, Baker N. The PROPKA web interface. 2006.
32. Forsyth WR, Gilson MK, Antosiewicz J, Jaren OR, Robertson AD. Theoretical and experimental analysis of ionization equilibria in ovomucoid third domain. Biochemistry 1998;37:8643–8652.
33. Mauguen Y, Hartley RW, Dodson EJ, Dodsen GG, Bricogne G, Chothia C, Jack A. Molecular structure of a new family of ribonucleases. Nature 1982;297:162–164.
34. Dunback RL, Jr, Cohen FE. Baysian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6:1661–1681.
35. Bartik K, Redfield C, Dobson CM. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional 1H NMR. Biophys J 1994;66:1180–1184.
36. Kuramitsu S, Hamaguchi K. Analysis of the acid-base titration curve of hen lysozyme. J Biochem 1980;87:1215–1219.
37. White A, Rose DR. Mechanism of catalysis by retaining β-glycosyl hydrolases. Curr Opin Struct Biol 1997;7:645–651.
38. Joshi MD, Hedberg A, McIntosh LP. Complete measurement of the pKa values of the carboxyl and imidazole groups in *Bacillus circulans* xylanase. Protein Sci 1997;6:2667–2670.
39. McIntosh LP, Hand G, Johnson PE, Joshi MD, Körner M, Plesniak LA, Ziser L, Wakarchuk WW, Withers SG. The pKa of the general acid/base carboxyl group of a glycosidase cycles during catalysis: a ^{13}C-NMR study of *Bacillus circulans* xylanase. Biochemistry 1996; 35:9958–9966.
40. Berisio R, Lamzin VS, Sica F, Wilson KS, Zagari A, Mazzarella L. Protein titration in the crystal state. J Mol Biol 1999;292:845–854.
41. Berisio R, Sica F, Lamzin VS, Wilson KS, Zagari A, Mazzarella L. Atomic resolution structures of ribonuclease A at six pH values. Acta Crystallogr 2002;D58:441–450.
42. Howlin B, Moss DS, Harris GW. Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. Acta Crystallogr 1989;A45:851–861.
43. Wlodawer A, Svensson LA, Sjolin L, Gilliland GL. Structure of phosphate-free ribonuclease A refined at 1.26 A. Biochemistry 1988; 27:2705–2717.
44. Mel VSJd, Doscher MS, Martin PD, Edwards BFP. The occupancy of two distinct conformations by active-site histidine-119 in crystals of ribonuclease is modulated by pH. FEBS 1994;349:155–160.
45. Yang W, Hendrickson WA, Crouch RJ, Satow Y. Structure of ribonuclease H phased at 2 A resolution by MAD analysis of the selenomethionyl protein. Science 1990;249:1398–1405.
46. Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Nakamura H, Ikehara M, Matsuzaki T, Morikawa K. Structural details of ribonuclease H from *Escherichia coli* as refined to an atomic resolution. J Mol Biol 1992;223:1029–1052.
47. Katayanagi K, Okumura M, Morikawa K. Crystal structure of *Escherichia coli* RNase HI in complex with Mg2+ at 2.8 A resolution: proof for a single Mg(2+)-binding site. Proteins: Struct Funct Genetics 1993;17:337–346.
48. Oda Y, Yamazaki T, Nagayama K, Kanaya S, Kuroda Y, Nakamura H. Individual ionization constants of all the carboxyl groups in ribonuclease HI from *Escherichia coli* determined by NMR. Biochemistry 1994;33:5275–5284.
49. Oliveberg M, Arcus VL, Fersht AR. pKa values of carboxyl groups in the native and denatured states of barnase: the pKa values of the denatured state are on average 0.4 units lower than those of model compounds. Biochemistry 1995;34:9424–9433.
50. Meadows DH, Roberts GCK, Jardetzky O. Nuclear magnetic resonance studies of the structure and binding sites of enzymes. VIII. Inhibitor binding to ribonuclease. J Mol Biol 1969;45:491–511.
51. Markley JL. Correlation proton magnetic resonance studies at 250 MHz of bovine pancreatic ribonuclease. I. reinvestigation of the histidine peak assignments. Biochemistry 1975;14:3546–3553.
52. Walters DE, Allerhand A. Tautomeric states of the histidine residues of bovine pancreatic ribonuclease A. Application of carbon 13 nuclear magnetic resonance spectroscopy. J Biol Chem 1980;255:6200–6204.
53. Markley JL. Correlation proton magnetic resonance studies at 250 MHz of bovine pancreatic ribonuclease. II. pH and inhibitor-induced conformational transitions affecting histidine-48 and one tyrosine residue of ribonuclease A. Biochemistry 1975;14:3554–3561.
54. Humphrey W, Dalke A, Schulten K. VMD—visual molecular dynamics. J Mol Graph Model 1996;14:33–38.