Protein Cell 2012, 3(3): 230–238
DOI 10.1007/s13238-012-2035-4

Protein & Cell

RESEARCH ARTICLE

Calculating pH-dependent free energy of proteins by using Monte Carlo protonation probabilities of ionizable residues

Qiang Huang[1] ✉, Andreas Herrmann[2]

[1] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China
[2] Department of Biology, Molecular Biophysics, Humboldt University of Berlin, 10115 Berlin, Germany
✉ Correspondence: huangqiang@fudan.edu.cn
Received December 21, 2011 Accepted January 9, 2012

ABSTRACT

Protein folding, stability, and function are usually influenced by pH. And free energy plays a fundamental role in analysis of such pH-dependent properties. Electrostatics-based theoretical framework using dielectric solvent continuum model and solving Poisson-Boltzmann equation numerically has been shown to be very successful in understanding the pH-dependent properties. However, in this approach the exact computation of pH-dependent free energy becomes impractical for proteins possessing more than several tens of ionizable sites (e.g. > 30), because exact evaluation of the partition function requires a summation over a vast number of possible protonation microstates. Here we present a method which computes the free energy using the average energy and the protonation probabilities of ionizable sites obtained by the well-established Monte Carlo sampling procedure. The key feature is to calculate the entropy by using the protonation probabilities. We used this method to examine a well-studied protein (lysozyme) and produced results which agree very well with the exact calculations. Applications to the optimum pH of maximal stability of proteins and protein–DNA interactions have also resulted in good agreement with experimental data. These examples recommend our method for application to the elucidation of the pH-dependent properties of proteins.

KEYWORDS protein protonation, protein electrostatics, pH-dependent free energy, Poisson-Boltzmann equation, Monte Carlo simulation

INTRODUCTION

Background

The structural stability and function of proteins are usually influenced by the concentration of hydrogen ions (pH). Changes in pH, for example of the cellular environment, shift the protonation equilibria of the ionizable (or titratable) residues of proteins, and consequently alter their charges and the electrostatic interactions among residues, resulting in a marked pH-dependence of protein folding, stability, and activity. Thus, mechanistic understanding of the pH-dependent properties of proteins is of strong interest.

In the past two decades, significant advances have been made in modeling of the pH-dependent properties of proteins (Antosiewicz et al., 1994; Bashford, 2004). Especially, the electrostatics-based framework using the dielectric solvent continuum model and solving the Poisson-Boltzmann (PB) equation numerically (i.e. the framework of pH-dependent continuum electrostatics) have been shown to be remarkably accurate for a wide range of applications, including average protonation of ionizable residues (Beroza et al., 1991), pK_a shifts of ionizable residues (Bashford and Karplus, 1991), pH-induced conformational changes of proteins (Huang et al., 2002; Langella et al., 2006), pH-dependence of enzymatic activity (Tynan-Connolly and Nielsen, 2007), and pH-dependent protein–protein (Dong et al., 2003), protein–DNA (or RNA) interactions (Misra et al., 1998; Olson, 2001; Frick et al., 2004), etc. Indeed, several web servers for the calculations of pH-dependent properties of proteins have become available (Gordon et al., 2005; Kantardjiev and Atanasov, 2006; Tynan-Connolly and Nielsen, 2006).

Theoretical treatment of pH-effects on the stability and ac-

tivity of a protein requires evaluation of pH-dependent free energy of the protein. The free energy is one of the most central quantities in biomolecular modeling. However, it is one of the most difficult quantities to compute on the basis of atomic-level simulations (Rodinger and Pomès, 2005; Meirovitch, 2007). In the theoretical framework of continuum electrostatics, the protonation microstate of a protein possessing *N* ionizable sites is characterized with an *N*-component vector, in which each component is a two-value variable defining the microstate of an ionizable site: for the protonated site occupied by a proton, the variable is set to 1; for the unprotonated site, to 0. Thus, evaluation of free energy involves the computation of the partition function of 2^N protonation microstates for the protein (see also Theory). To exactly compute all 2^N microstate energies is very time-consuming and, owing to the current capacities of computation, becomes impractical for the proteins with more than several tens of ionizable sites (e.g. 2^{30} = 1,073,741,824 microstates for 30 ionizable residues). Since the numbers of those sites of many proteins are significantly greater than 30, reliable methods to estimate the protonation partition function and, hence the free energy, are necessary for modeling the pH-dependence of protein properties.

Several methods for estimation of the protonation partition function have been developed. One of them is a rather direct method which sums up the low-energy protonation microstates obtained by a Monte Carlo sampling procedure based on Metropolis algorithm (Metropolis et al., 1953). However, a special procedure for comparison of the microstate energies has to be designed for collecting the lowest microstate energies and assuring that each microstate is taken only once (Alexov and Gunner, 1999; Alexov, 2004). Another method is an application of the proton linkage model which relates the pH-induced change in the relative free energy to the average number of protons released from the protein when shifting from a reference pH to a desired pH (Tanford, 1970; Yang and Honig, 1993; Misra et al., 1998).

In this study, we present a method to compute the ionization free energy directly using the average protonation energy and the protonation probabilities of ionizable sites estimated by a well-established Monte Carlo method (Beroza et al., 1991). As justified by theoretical considerations, the method treats the protonation entropy of protein as the sum of the entropy contributions from individual ionizable sites. Calculation of the entropy is based on the protonation probabilities of the ionizable sites which can be obtained by the Monte Carlo procedure (Beroza et al., 1991). Testing the accuracy of this method by examining a well-studied protein, the lysozyme, produced results which agree very well with the exact calculations. To further demonstrate the potential, the method was applied to calculate the optimum pH of maximal stability of proteins and the pH-dependent free energy of protein–DNA interaction, and the results show a good agreement with the experimental data. These examples as shown in Fig. 1 demonstrated that our method is capable of evaluating the pH-dependent free energy of proteins.

Theory

We employ the modeling framework of protein protonation using the dielectric continuum solvent model (Ullmann and Knapp, 1999). This framework uses an *N*-component vector, $\mathbf{X}=(x_1,\ldots,x_i,\ldots,x_N)$, to define the protonation microstate of a protein possessing *N* ionizable sites. The vector component, x_j, is a two-value variable that defines the protonation microstate of ionizable site *j*: for the protonated site occupied by a proton, x_j is set to 1; for the unprotonated site, x_j is set to 0. Taking the neutral protonation state (i.e. acids are protonated, and bases are unprotonated) as the reference (designated as $\mathbf{X}^0$), the potential energy of the protein in a protonation microstate (designated as $\mathbf{X}^n$) at the given pH and temperature *T* is

$$E_n = \sum_{j=1}^{N}\left(x_j^n - x_j^0\right)k_{\mathrm{B}}T\ln 10\left(\mathrm{pH}-\mathrm{p}K_{a,j}^{intr}\right) + \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(x_j^n + z_j^0\right)\left(x_k^n + z_k^0\right)B_{jk}, \quad (1)$$

where k_{B} is the Boltzmann constant, and the quantities of the reference state are indicated by the superscript "0": x_j^0 is 1 for acids and 0 for bases, $\mathrm{p}K_{a,j}^{intr}$ is the intrinsic $\mathrm{p}K_a$ value of site *j*, z_j^0 is the formal charge of site *j* in the unprotonated state (i.e. –1 for acids and 0 for bases), and B_{jk} is the electrostatic interaction between two unit charges at sites *j* and *k*. Therefore, the partition function of protein protonation is

$$Z = \sum_{n}^{2^N} e^{-\beta E_n}, \quad (2)$$

where $\beta = 1/(k_{\mathrm{B}}T)$. In principle, the ionization free energy from the reference (neutral protonation) state at the given pH and temperature *T* can be directly calculated according to $\Delta G=-k_{\mathrm{B}}T\ln Z$. As mentioned, however, it becomes computationally impractical for proteins with more than several tens of ionizable sites (e.g. > 30), because the computation requires a summation over all possible 2^N protonation microstates of the protein. For such proteins, one has to use an approximation method.

In order to develop an approximation method, we focus on the ensemble averages of the potential energy and the protonation probabilities of ionizable sites, which are given, respectively, by

$$U = \langle E_n \rangle = \frac{\sum_n^{2^N} E_n e^{-\beta E_n}}{Z}, \quad (3)$$

and

$$\theta_j = \langle x_j \rangle = \frac{\sum_n^{2^N} x_j e^{-\beta E_n}}{Z} (j = 1,2,\ldots,N). \quad (4)$$

Since each ionizable site has only two possible protona-

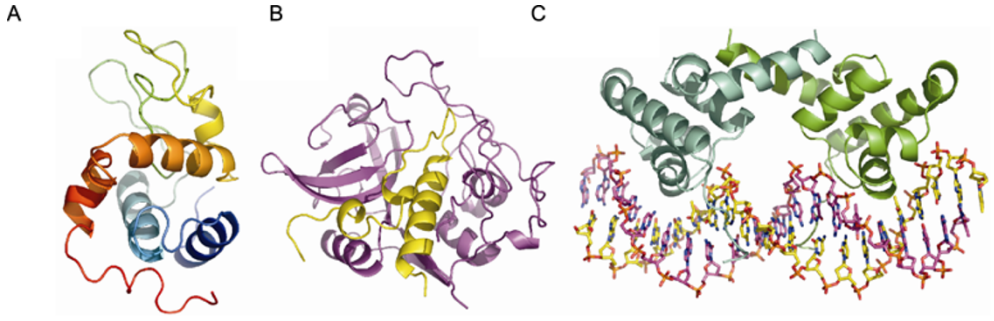

Figure 1. The structures used in the calculations of this study. (A) lysozyme (PDB code: 7LYZ), where the protein chain is colored with rainbow spectrum. (B) Cathepsin B (PDB code: 1HUC), where chain A is in yellow, and chain B in purple. (C) λcI repressor-operator complex (PDB code: 1LMB), where two protein chains are in greencyan and green, respectively, and two DNA chains in purple and yellow, respectively.

tion microstates (i.e. protonated: $x_j = 1$, and unprotonated: $x_j = 0$), the average probability of site j to be in the "unprotonated" state is

$$\lambda_j = 1 - \theta_j . \quad (5)$$

According to the Gibbs formula for the entropy, with the protonation and "unprotonation" probabilities, the protonation entropy of site j is

$$s_j = -k_B(\theta_j \ln \theta_j + \lambda_j \ln \lambda_j) . \quad (6)$$

On the other hand, the overall protonation entropy of a protein is usually given by

$$S = -k_B \sum_n^{2^N} p_n \ln p_n , \quad (7)$$

where p_n is the probability of finding the protein in the microstate, $\mathbf{X}^n$, and given by

$$p_n = \frac{e^{-\beta E_n}}{Z} . \quad (8)$$

As we will show in Appendix A, Eq. 7 can theoretically be expressed as the sum of the (entropy) contributions from the N individual sites:

$$S = \sum_j^N s_j . \quad (9)$$

This relationship leads to a more direct method to estimate the free energy. To avoid the summation over all 2^N possible microstates, we employ the well-established Monte Carlo sampling procedure (Beroza et al., 1991) to obtain (i.e. to approximate) the average protonation energy U and the protonation probabilities, $\theta_j (j = 1,2,\ldots,N)$, according to Eqs. 3 and 4, and thereby the entropy with Eqs. 5, 6, and 9. Since the pressure-volume term is negligible, the (Gibbs) ionization free energy is then approximated by the average energy and the entropy, as

$$G^{pH} \approx U_{mc} - TS_{mc} , \quad (10)$$

where the subscript '*mc*' indicates the values estimated by the Monte Carlo method. Details about the Monte Carlo procedures for evaluating the average protonation energy and the protonation probabilities in Eq. 10 can also be referred to these studies (Beroza et al., 1991; Ullmann and Knapp, 1999; Rabenstein and Knapp, 2001). Note that, as mentioned, the ionization free energy in Eq. 10 is a relative value from the reference state, i.e. the neutral protonation state.

RESULTS

Comparison with the exact calculations for lysozyme

We first tested the accuracy of the method by examining a small, well-studied protein, lysozyme, in which only 21 of the 32 ionizable sites are relevant for titration, due to that 11 arginines have very little effects on other residues titration in the pH range from 0 to 12 (Beroza et al., 1991). We calculated the exact ionization free energies of the lysozyme in the pH range from 0 to 12 with the partition function requiring the sum over all 2^{21} protonation microstates (Eq. 2). Alternatively, the ionization free energy was estimated by the average energy and entropy obtained by 10,000 Monte Carlo sampling steps, respectively. The intrinsic pK_a values and site–site interactions used here were the same as those used in prior studies (Bashford and Karplus, 1990; Beroza et al., 1991). Details can be seen in the calculation example included in the MEAD suite of programs.

Figure 2A shows that the exact free energies and those obtained by 10,000 Monte Carlo sampling steps are almost the same. For further comparison of both approaches, we calculated the exact values of the average energy and of the entropy according to Eqs. 3 and 7 by a sum over all 2^{21} microstates. The values from both methods are almost the same (Fig. 2B and 2C). In particular, the Monte Carlo-based values of the entropy are almost identical to those by the exact calculations (Fig. 2C), demonstrating that Eqs. 7 and 9 are equivalent, just as theoretically justified in Appendix A. Also, the results indicate that the efficient Monte Carlo sampling procedure (Beroza et al., 1991) provides a solid basis for our method.

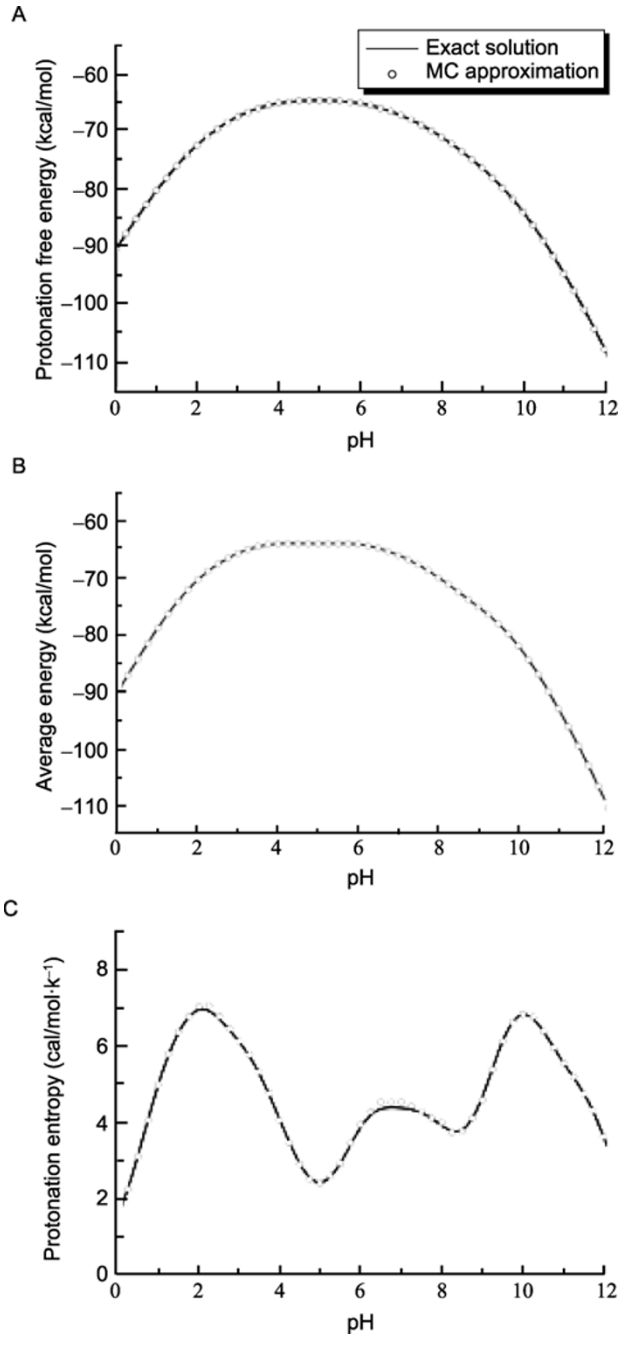

Figure 2. The ionization free energies of the lysozyme in pH range 0–12 obtained by the exact calculation and our method. (A) The ionization free energy. (B) The average energy. (C) The protonation entropy.

Optimum pH of maximal stability of Cathepsin B

Proteins are known to differ in the optimum pH of their stability. Numerical calculations of pH-dependent free energy would support the understanding of the molecular origin of the specific pH of protein stability, and may also provide guidelines for the design of proteins which are stable and active in a certain range of pH (Tynan-Connolly and Nielsen, 2006; Tynan-Connolly and Nielsen, 2007). In previous numerical calculations, the optimum pH was determined as the pH of the minimum free energy of folding. Alexov has calculated the optimum pHs for a number of proteins, for example Cathepsin B, using the Monte Carlo approximation method, with a special procedure that collects the lowest microstate energies and thus assures that each microstate is taken only once (Alexov, 2004).

Figure 3 shows the pH-dependent free energies of Cathepsin B in the pH range of 0–14 calculated with our approach. Three energies were computed: the free energy of the unfolded state, the free energy of the folded state, and the free energy of folding. The pH-dependent free energy of the unfolded state was calculated with the same method used by Alexov (Alexov, 2004). Different from Alexov's study, here the free energies of the folded state and folding were not scaled by an additive constant. The results indicate that the optimum pH of Cathepsin B is 5.15, in excellent agreement with the experimental value (Alexov, 2004). The shape of three curves of the pH-dependent free energies are very similar to those in the corresponding figure in Ref (Alexov, 2004), further demonstrating that our method obtained equivalent results, and is capable of calculating the pH-dependent free energy of proteins in the folded state.

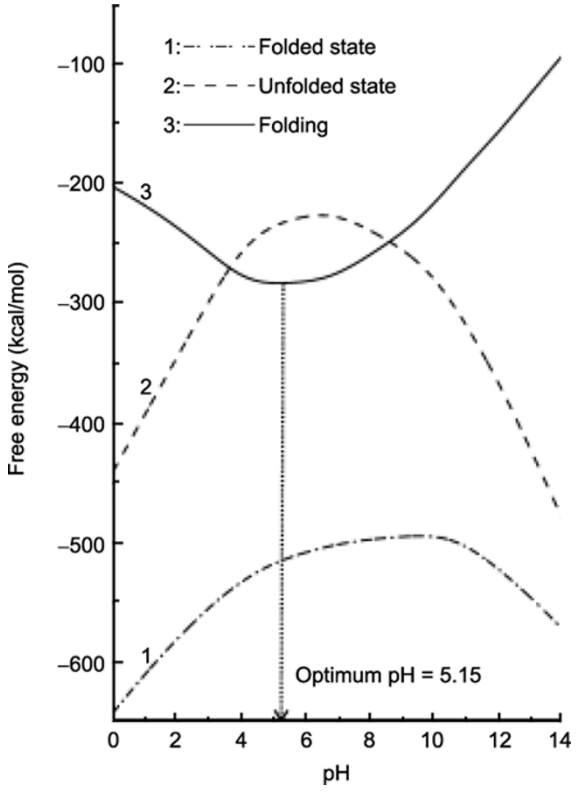

Figure 3. The pH-dependent contributions to the folded, unfolded, and folding free energies of Cathepsin B in pH range 0–14.

The pH-dependence of protein-DNA binding

Protein–DNA interactions are important for transcriptional regulation in both prokaryotes and eukaryotes. Because of the highly charged nature of nucleic acids, electrostatic forces and therefore pH play a significant role in protein–DNA interactions, e.g. the λcI repressor-operator system. It has been shown that the affinity of the repressor to its operators is highly dependent on pH, and pH effects contribute to the discrimination of DNA binding sites by the repressor (Senear and Ackers, 1990; Senear and Batey, 1991). Hence, to understand how a gene regulatory protein binds to specific DNA sequences, it is necessary to quantify the pH effects on the binding free energy of the protein-DNA system. In the past, the pH-dependent protein-DNA binding in the λcI repressor-operator system has been studied intensively for testing the accuracy of theoretical models used to calculate electrostatic free energies of protein-DNA binding. Misra et al. (1998) have used nonlinear Poisson-Blotzmann (NLPB) equation to elucidate the pH-dependence of the repressor-operator binding free energy. The calculated values were found to be in good agreement with the experimental data (Senear and Ackers, 1990). In their study, they employed an approximation method in which the pH effect on the binding was described with the expression given by Tanford (1970), who treated the problem of pH-dependence in terms of multiple equilibria involving acids and bases. To compare with this method, we computed the pH-dependent contribution to the binding free energy for this protein-DNA system by our approach.

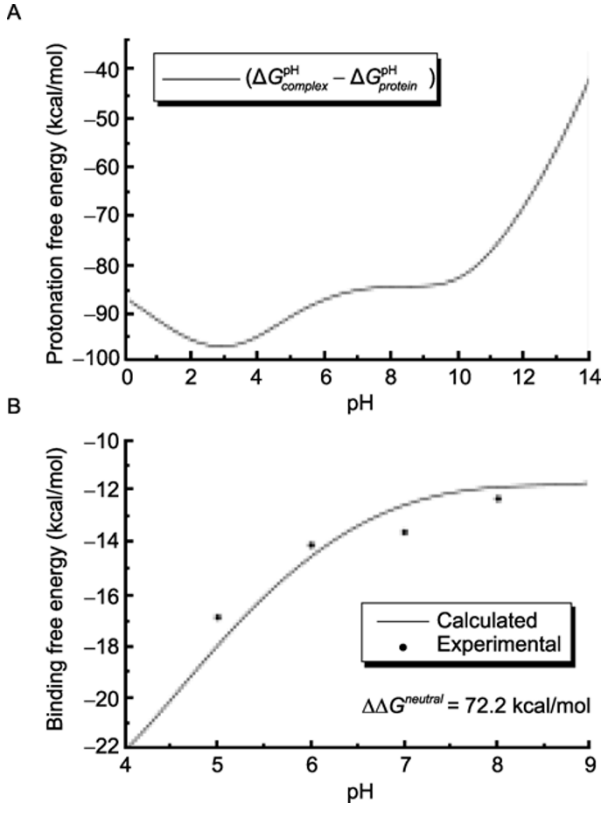

Figure 4. The pH-dependence of binding free energy of the λcI repressor-operator. (A) The change of ionization free energy upon the binding of the unbound λcI repressor to DNA in pH range 0–14. (B) The pH-dependent binding free energy in pH range 4–9 obtained with a best fitting pH-independent term of 72.7 kcal/mol.

Because the DNA titration has no significant role in the proton-linked effects at physiological pH (i.e. is independent of pH), the pH-dependent contribution to the binding free energy depends mainly on the ionization free energies of the protein-DNA complex and the unbound protein. Therefore, the binding free energy of the λcI repressor-operator system can be written as

$$\Delta\Delta G_{binding} = \Delta G^{pH}_{complex} - \Delta G^{pH}_{protein} + \Delta\Delta G^{neutral}, \quad (11)$$

where $\Delta G^{pH}_{complex}$ and $\Delta G^{pH}_{protein}$ are the ionization free energies of the complex and the unbound protein at the given pH, respectively, with respect to the hypothetical state in which all ionizable residues of the protein are neutral (i.e. the neutral protonation state). The additive constant term, $\Delta\Delta G^{neutral}$, is the pH-independent electrostatic contribution to the binding from the complex and the unbound protein in the neutral protonation state. To calculate the protonation energies, we first carried out MD simulations to generate conformational ensembles for the complex and the unbound protein, as described in MATERIALS AND METHODS. Fifty MD conformational snapshots of the complex or the unbound protein were extracted from the production run trajectories at the intervals of 20 ps. Then, the protonation energies of the snapshots at 20°C were calculated and averaged.

Figure 4A shows the difference in the protonation energy between the complex and the unbound protein, i.e. the pH-dependent contribution to the binding free energy. With a best-fitting pH-independent term of 72.7 kcal/mol (i.e. the $\Delta\Delta G^{neutral}$ value), the pH-dependent binding free energy was computed, as shown in Fig. 4B. The results illustrate that the pH-dependence of the calculated values is in good agreement with the experimental data (Senear and Ackers, 1990) and demonstrates the accuracy of our method in describing pH-dependent protein-DNA interactions. Note that the $\Delta\Delta G^{neutral}$ value does not influence the pH-dependence, but only shifts the binding free energy by the same value at each pH. However, in order to use the general expression given by Tanford (1970), Misra et al. have actually shown that the total electrostatic free energy of the repressor–operator interaction opposing binding is about 73 kcal/mol (Misra et al., 1998), which is very similar to our fitting value. Details about this and how single ionizable residues affect the binding free energy can be referred to the study by Misra et al. (1998).

DISCUSSION

Since the pH-dependent free energy plays a fundamental role

in the analysis of the pH-dependence of protein folding, stability and activity, accurate calculation of the pH-dependent free energy is one of the most important and challenging areas in the field of protein modeling and simulations. In the past, the framework of the pH-dependent continuum electrostatics has been proven to be a suitable basis for further methodology development. However, because the number of ionizable residues in most of proteins is more than several tens, e.g. 30, one has to use an approximation method. Previously, two methods based on Monte Carlo simulation have been developed (Misra et al., 1998; Alexov, 2004). Inspired by them, in this study we have developed a new Monte Carlo-based method. This method was tested and validated by comparisons with the exact calculations for lysozyme, with the experimental values of optimum pH of maximal stability of Cathepsin B and of the pH-dependence of the λcI repressor-operator binding.

Both Cathepsin B and the λcI repressor-operator binding have also been studied by the two previously developed Monte Carlo simulation methods (Misra et al., 1998; Alexov, 2004). The method by Alexov calculates the free energy with a direct estimation of the partition function by summing up the low-energy protonation microstate obtained by Monte Carlo sampling. As a result, this method needs a special comparison procedure to ensure that each microstate is taken only once (Alexov, 2004). In contrast, our method does not need such a procedure because it is not required for Monte Carlo-based determination of ensemble averages of the protonation energy and of the probabilities of ionizable sites. On the other hand, another method based on the proton linkage model provides an effective method to obtain the pH-dependent free energy, by relating the pH-induced change of reaction free energy to the average number of protons released from the considered molecular system (Tanford, 1970; Yang and Honig, 1993; Misra et al., 1998). Thus, like ours, this method also uses the Monte Carlo-based values of the protonation probabilities of ionizable sites. But a major difference is that our method does not need to define a reference pH for calculations.

As mentioned, usually an exact calculation of the ionization free energy is feasible only for those proteins with ionizable sites less than 30. Of course, in practice many sites of a protein with a large number of ionizable groups may be either fully protonated or unprotonated and thereby only a limited number of sites have fractional protonation probabilities that really contribute to the protonation entropy in Eq. 6. However, to determine which sites are fully protonated or unprotonated, one has to carry out Monte Carlo calculations in advance. Thus, approximation methods are needed for a large number of proteins. For this purpose, Monte Carlo approximation methods based on the Metropolis importance-sampling strategy have been developed, including ours. As shown in Theory section, the efficiency of the Monte Carlo sampling is crucial for the accuracy of our method. Fortunately, the Monte Carlo sampling procedure used for calculating the protonation probabilities has already been proven very effective, and has wide applications (Beroza et al., 1991; Kannt et al., 1998; Da Silva et al., 2001; Rabenstein and Knapp, 2001; Seiffert et al., 2007; Tynan-Connolly and Nielsen, 2007). It was found that about 10,000 Monte Carlo steps provide average protonation values with an absolute error of ≈ 0.02 protons (Beroza et al., 1991). Even for proteins with several hundreds of ionizable sites, the time required for sampling is reasonable, such as in our previous study on influenza virus hemagglutinin (Huang et al., 2002). Beroza et al. have pointed out that quantitative discrepancies between theory and experiment are attributed to the accuracy of the input to the Monte Carlo sampling, i.e. the electrostatic interactions between two unit charges at the ionizable sites in Eq. 1, rather than to the Monte Carlo method itself (Beroza et al., 1991). Therefore, the accuracy in the Monte Carlo sampling provides a solid basis for our method which uses the Monte Carlo-based protonation probabilities to calculate the protonation entropy.

In conclusion, we have presented a method for calculating pH-dependent free energy of proteins using the dielectric solvent continuum model and solving the Poisson-Boltzmann equation numerically. Our method computes the free energy directly using the average energy and the protonation probabilities of ionizable sites obtained by the well-established Monte Carlo sampling procedure. The key feature of this method is to calculate the protonation entropy by using the Monte Carlo-based protonation probabilities. Thus, the method can be implemented easily, and its application results here support that this method may find potential application to the elucidation of the pH-dependent properties of proteins.

MATERIALS AND METHODS

Structural models and partial charges of atoms

The ionization free energies of lysozyme, Cathepsin B, barnase-barstar binding, and λcI repressor-operator were calculated with the described method (Fig. 1). The structural models for the lysozyme and corresponding parameters such as the partial charges of atoms were taken directly from the calculation example included in the MEAD (Macroscopic Electrostatics with Atomic Detail) suite of programs (Bashford and Gerwert, 1992; Bashford, 1997). These data have been used in previous studies by Bashford and Karplus (1990), and Beroza et al. (1991). The structures for Cathepsin B and the λcI repressor-operator were generated by MD simulation described below. For Cathepsin B, an MD simulation system was constructed, and the starting coordinates of the protein are those from the 2.15 Å resolution crystal structure in the Protein Data Bank (PDB code: 1HUC) (Musil et al., 1991). For the λcI repressor-operator, two MD simulation systems were constructed for the unbound protein (i.e. the λcI repressor) and the protein-DNA (i.e. λcI repressor-operator) complex, respectively; the starting coordinates of the unbound protein and the complex were taken from the refined 1.8 Å resolution crystal

structure (PDB code: 1LMB) (Beamer and Pabo, 1992). For MD simulations, CHARMM27 force field (MacKerell et al., 1998) was used for all three systems. Thus, the partial charges of atoms used for calculating the ionization free energy were also assigned according to the CHARMM27 force field. To compute the ionization free energy, for each system (Cathepsin B, barnase, barstar, barnase-barstar complex, the λcI repressor, the λcI repressor-operator complex) we extracted 50 conformational snapshots of the structure from the production MD simulation runs, as described in the following subsection.

Molecular dynamics simulations

MD simulations were carried out to generate conformational ensembles of Cathepsin B, barnase, barstar, the λcI repressor, and the λcI repressor- operator complex for the calculations of ionization free energies (for starting structures see above). To construct a simulation system, the starting all-hydrogen structure of Cathepsin B (or λcI repressor, or λcI repressor-operator complex) was merged into a rectangular box of TIP3P (Jorgensen et al., 1983) water molecules. The thickness of the water layer between the protein and the closest box-boundary was ~14 Å. The ionic concentrations in the water box were set to those as used in corresponding experimental studies, i.e. for Cathepsin B 100 mmol/L, and for the λcI repressor and the λcI repressor-operator complex, 200 mmol/L KCl. Equal numbers of Na^+ (or K^-) and Cl^- ions were first determined according to the given ionic concentrations, and then additional Cl^- ions were placed into the box to neutralize the system. Simulations were run under *N, P, T* conditions using the program NAMD (Kale et al., 1999), with the pressure *P* = 1 atm, and the temperature *T* = 300 K (for Cathepsin B) or 293 K (for two λcI repressor-operator systems). Periodic boundary conditions, particle-mesh Ewald (PME) method of the electrostatic forces (Essmann et al., 1995), SHAKE algorithm (Ryckaert et al., 1977), and 1-fs time step were employed. The period for each simulation run was 2 ns: 1-ns equilibration phase was first completed, and then 1-ns production phase was collected. To calculate the ionization free energy, 50 MD conformational snapshots of Cathepsin B (or barnase-barstar systems, or λcI repressor-operator systems) were extracted from the 1-ns production run at time intervals of 20 ps.

Calculation of ionization free energy

The MEAD suite of programs (Bashford and Gerwert, 1992; Bashford, 1997) was employed for electrostatic calculations based on the continuum solvent model. With the atomic coordinates of every MD conformational snapshot, the program *multiflex* was used to compute the electrostatic interactions of charged ionizable residue pairs, B_{jk}, in Eq. 1, with the dielectric constants $\varepsilon_p = 4$ inside the protein and $\varepsilon_S = 80$ for the solvent. For the computation of the electrostatic potential, the finite difference lattice was set up as a cubic box containing the whole structure. Electrostatic potentials were calculated with two focusing steps: a grid spacing of 1.0 Å (grid centered at the protein) and of 0.25 Å (grid centered at the ionizable group). With the obtained $pK_{a,j}^{intr}$ and B_{jk}, the Monte Carlo procedure implemented in the program *Karlsberg* (Rabenstein and Knapp, 2001) was then used to calculate the average energy *U* in Eq. 3 and the protonation probability, Θ_i, in Eq. 4. To compute the probability of an individual ionizable site, 10,000 protonation microstates were sampled with a standard deviation less than 0.01 protons for each ionizable site. Then, the ionization free energy was calculated based on Eqs 5, 6, and 9. For the calculations using the MD snapshots, the free energy was first calculated for every snapshot and then averaged over all 50 snapshots.

ACKNOWLEDGEMENTS

We thank Prof. Dr. Ernst-Walter Knapp (Free University of Berlin, Germany) for his comments in the initial phase of this study. This study was supported in part by grants from the National Natural Science Foundation of China (Grant No. 30570406), the HI-tech Research and Development Program of China (Grant No. 2008AA02Z311), and the Shanghai Leading Academic Discipline Project (B111).

Appendix A

Considering a statistical-mechanical system that consists of *N* distinguished components (e.g. particles, or, ionizable sites of protein in our case), a microstate of the system is determined by a specific set of the one-dimensional "coordinates" of the components, i.e. the state vector, $\mathbf{x} = (x_1,\ldots,x_j,\ldots,x_N)$. To derive Eq. 7, we begin with a state vector **X** with *N* continuous components, $x_j \in [0,1] (j = 1,2,\ldots,N)$. Accordingly, the probability of finding the system in a specific microstate, **X**, is

$$p(\mathbf{x}) = \frac{e^{-\beta E(\mathbf{x})}}{\int_{\mathbf{x}} e^{-\beta E(\mathbf{x})} d\mathbf{x}}, \tag{A1}$$

where $E(\mathbf{X})$ is the potential energy of the microstate. According to the Gibbs formula for the entropy, the (configurational) entropy of the system is given by

$$S = -k_B \int_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \tag{A2}$$

On the other hand, for component *j*, one may define the probability of finding the component in the microstate x_i as $p(x_i)$, which is satisfied with

$$\int_0^1 p(x_j) dx_j = 1. \tag{A3}$$

Because the so-called microstate of the system, **X** is actually the combination of the microstates of the *N* individual components, and the combined probability of the *N* components to be in the microstate **X** is the product of those of the individual components, we have

$$p(\mathbf{x}) = \prod_{j=1}^{N} p(x_j). \tag{A4}$$

Substitute Eq. A4 into A2 and consider Eq. A3, one can show

$$S = -k_B \sum_{j=1}^{N} \int_0^1 p(x_j) \ln p(x_j) dx_j. \tag{A5}$$

Specific to the protein protonation, the component x_i is a two-value variable (1 or 0), and the averages are protonation probability $p(x_j = 1) = \theta_j$ and unprotonation probability $p(x_j = 0) = \lambda_j$ (see Eqs. 4 and 5), respectively. So, based on the general Eqs. A2 and A5, we have

$$S = -k_B \sum_{n=1}^{2^N} p_n \ln p_n = -k_B \sum_{j=1}^{N} \left(\theta_j \ln \theta_j + \lambda_j \ln \lambda_j \right). \quad \text{(A6)}$$

Therefore, the protonation entropy of the protein is the sum of the contributions from *N* individual ionizable sites, as shown in Eq. 7.

REFERENCES

Alexov, E. (2004). Numerical calculations of the pH of maximal protein stability. The effect of the sequence composition and three-dimensional structure. Eur J Biochem 271, 173–185.

Alexov, E.G., and Gunner, M.R. (1999). Calculated protein and proton motions coupled to electron transfer: electron transfer from QA- to QB in bacterial photosynthetic reaction centers. Biochemistry 38, 8253–8270.

Antosiewicz, J., McCammon, J.A., and Gilson, M.K. (1994). Prediction of pH-dependent properties of proteins. J Mol Biol 238, 415–436.

Bashford, D. (1997). An object-oriented programming suite for electrostatic effects in biological molecules. In: Scientific computing in object-oriented parallel environments. Y. Ishikawa, R.R. Oldehoeft, J.V.W. Reynders, and M. Tholburn, eds. Berlin: Springer. 233–240.

Bashford, D. (2004). Macroscopic electrostatic models for protonation states in proteins. Front Biosci 9, 1082–1099.

Bashford, D., and Gerwert, K. (1992). Electrostatic calculations of the pKa values of ionizable groups in bacteriorhodopsin. J Mol Biol 224, 473–486.

Bashford, D., and Karplus, M. (1990). pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 29, 10219–10225.

Bashford, D., and Karplus, M. (1991). Multi-site titration curves of proteins: An analysis of exact and approximate methods for their calculation. J Chem Phys 95, 9556–9561.

Beamer, L.J., and Pabo, C.O. (1992). Refined 1.8 A crystal structure of the lambda repressor-operator complex. J Mol Biol 227, 177–196.

Beroza, P., Fredkin, D.R., Okamura, M.Y., and Feher, G. (1991). Protonation of interacting residues in a protein by a Monte Carlo method: application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. Proc Natl Acad Sci U S A 88, 5804–5808.

Da Silva, F.L.B., Jönsson, B., and Penfold, R. (2001). A critical investigation of the Tanford-Kirkwood scheme by means of Monte Carlo simulations. Protein Sci 10, 1415–1425.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. J Chem Phys 103, 8577–8593.

Frick, D.N., Rypma, R.S., Lam, A.M.I., and Frenz, C.M. (2004). Electrostatic analysis of the hepatitis C virus NS3 helicase reveals both active and allosteric site locations. Nucleic Acids Res 32, 5519–5528.

Gordon, J.C., Myers, J.B., Folta, T., Shoja, V., Heath, L.S., and Onufriev, A. (2005). H++: a server for estimating pKas and adding missing hydrogens to macromolecules. Nucleic Acids Res 33, W368–371.

Huang, Q., Opitz, R., Knapp, E.W., and Herrmann, A. (2002). Protonation and stability of the globular domain of influenza virus hemagglutinin. Biophys J 82, 1050–1058.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J Chem Phys 79, 926–935.

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K., and Schulten, K. (1999). NAMD2: Greater scalability for parallel molecular dynamics. J Comput Phys 151, 283–312.

Kannt, A., Lancaster, C.R.D., and Michel, H. (1998). The coupling of electron transfer and proton translocation: electrostatic calculations on Paracoccus denitrificans cytochrome c oxidase. Biophys J 74, 708–721.

Kantardjiev, A.A., and Atanasov, B.P. (2006). PHEPS: web-based pH-dependent protein electrostatics server. Nucleic Acids Res 34, W43–47.

Langella, E., Improta, R., Crescenzi, O., and Barone, V. (2006). Assessing the acid-base and conformational properties of histidine residues in human prion protein (125-228) by means of pKa calculations and molecular dynamics simulations. Proteins: Struct Funct Bioinfo 64, 167–177.

MacKerell, A.D.J., Bashford, D., Bellot, M., Dunbrack, R.L.J., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem 102, 3586–3616.

Meirovitch, H. (2007). Recent developments in methodologies for calculating the entropy and free energy of biological systems by computer simulation. Curr Opin Struct Biol 17, 181–186.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equation of state calculations by fast computing machines. J Chem Phys 21, 1087–1092.

Misra, V.K., Hecht, J.L., Yang, A.-S., and Honig, B. (1998). Electrostatic contributions to the binding free energy of the lambdacI repressor to DNA. Biophys J 75, 2262–2273.

Musil, D., Zucic, D., Turk, D., Engh, R.A., Mayr, I., Huber, R., Popovic, T., Turk, V., Towatari, T., Katunuma, N., et al. (1991). The refined 2.15 A X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. EMBO J 10, 2321–2330.

Olson, M.A. (2001). Calculations of free-energy contributions to protein-RNA complex stabilization. Biophys J 81, 1841–1853.

Rabenstein, B., and Knapp, E.W. (2001). Calculated pH-dependent population of CO-myoglobin conformers. Biophys J 80, 1141–1150.

Rodinger, T., and Pomès, R. (2005). Enhancing the accuracy, the efficiency and the scope of free energy simulations. Curr Opin Struct Biol 15, 164–170.

Ryckaert, J.P., Ciccitti, G., and Berendsen, H.J.C. (1977). Numerical integration of the cartesian equations of motion of a system with constrains: molecular dynamics of n-alkanes. J Chem Phys 23, 327–341.

Seiffert, G.B., Ullmann, G.M., Messerschmidt, A., Schink, B., Kroneck,

P.M.H., and Einsle, O. (2007). Structure of the non-redox-active tungsten/[4Fe:4S] enzyme acetylene hydratase. Proc Natl Acad Sci U S A 104, 3073–3077.

Senear, D.F., and Ackers, G.K. (1990). Proton-linked contributions to site-specific interactions of lambda cI repressor and OR. Biochemistry 29, 6568–6577.

Senear, D.F., and Batey, R. (1991). Comparison of operator-specific and nonspecific DNA binding of the lambda cI repressor: [KCl] and pH effects. Biochemistry 30, 6677–6688.

Tanford, C. (1970). Protein denaturation. C. Theoretical models for the mechanism of denaturation. Adv Protein Chem 24, 1–95.

Tynan-Connolly, B.M., and Nielsen, J.E. (2006). pKD: re-designing protein pKa values. Nucleic Acids Res 34, W48-51.

Tynan-Connolly, B.M., and Nielsen, J.E. (2007). Redesigning protein pKa values. Protein Sci 16, 239–249.

Ullmann, G.M., and Knapp, E.W. (1999). Electrostatic models for computing protonation and redox equilibria in proteins. Eur Biophys J 28, 533–551.

Yang, A.-S., and Honig, B. (1993). On the pH dependence of protein stability. J Mol Biol 231, 459–474.

