A fast and accurate computational approach to protein ionization

VELIN Z. SPASSOV AND LISA YAN
Accelrys Inc., San Diego, California 92121, USA
(RECEIVED May 7, 2008; FINAL REVISION August 12, 2008; ACCEPTED August 15, 2008)

Abstract

We report a very fast and accurate physics-based method to calculate pH-dependent electrostatic effects in protein molecules and to predict the pK values of individual sites of titration. In addition, a CHARMm-based algorithm is included to construct and refine the spatial coordinates of all hydrogen atoms at a given pH. The present method combines electrostatic energy calculations based on the Generalized Born approximation with an iterative mobile clustering approach to calculate the equilibria of proton binding to multiple titration sites in protein molecules. The use of the GBIM (Generalized Born with Implicit Membrane) CHARMm module makes it possible to model not only water-soluble proteins but membrane proteins as well. The method includes a novel algorithm for preliminary refinement of hydrogen coordinates. Another difference from existing approaches is that, instead of monopeptides, a set of relaxed pentapeptide structures are used as model compounds. Tests on a set of 24 proteins demonstrate the high accuracy of the method. On average, the RMSD between predicted and experimental pK values is close to 0.5 pK units on this data set, and the accuracy is achieved at very low computational cost. The pH-dependent assignment of hydrogen atoms also shows very good agreement with protonation states and hydrogen-bond network observed in neutron-diffraction structures. The method is implemented as a computational protocol in Accelrys Discovery Studio and provides a fast and easy way to study the effect of pH on many important mechanisms such as enzyme catalysis, ligand binding, protein–protein interactions, and protein stability.

Keywords: protein ionization; pK prediction; continuum electrostatics; Generalized Born; CHARMm; hydrogen-bond network

The development of accurate methods to study ionization processes in proteins is important because pH-dependent changes in the protonation state can affect almost all molecular mechanisms related to protein function and stability. The modeling of protein ionization has a very long history. After the pioneering work of Linderstrom-Lang (1924), Tanford and Kirkwood (1957), and Tanford and Roxby (1972), a large number of methods have been proposed to model the titration of acidic and basic residues in protein molecules. The majority of contemporary physical models combine statistical thermodynamics with continuum electrostatics (Bashford and Karplus 1990, 1991). Most of these methods are based on finite-differences techniques to solve the Poisson–Boltzmann equation (FDPB) (Yang et al. 1993; Antosiewicz et al. 1994; Alexov and Gunner 1997; Schaefer et al. 1997). Some FDPB-based models have been extended to protein-membrane systems by including an implicit membrane in the solvation models (Karshikoff et al. 1994; Engels et al. 1995). A detailed description of the physical formalism of the electrostatic approach (Bashford 2004) and reviews of existing methods can be found in the literature (Honig and Nicholls 1995; Juffer 1998; Simonson 2001; Fitch and García-Moreno 2006). However, despite the progress in theoretical understanding of the problem, accurate prediction of ionization

Reprint requests to: Velin Z. Spassov, Accelrys Inc., 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA; e-mail: vss@accelrys.com; fax: (858) 799-5100.
Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.036335.108.

properties of proteins remains a challenge in protein modeling (Simonson 2001). The main sources of errors are well known, including the imperfectness in experimental structures used as input data, the use of crystal structures as models of solvated proteins, and the simplification of nonuniform dielectric properties of protein interior. In addition, due to the inherent coupling between the binding of protons and conformational changes (Laskowski Jr. and Sheraga 1954), flexibility of protein structures creates a serious combinatorial problem if the combined states of protonation and conformation are treated explicitly (Spassov and Bashford 1999). The easiest way to reduce the combinatorial problem is to ignore the conformational flexibility completely as in most popular FDPB and semi-empirical models. In contrast, traditional methods of molecular mechanics and molecular dynamics are typically based on the exact opposite approximation, completely ignoring the possible changes of protonation states. However, ignoring protonation changes can be misleading in many interesting cases such as predicting the properties of active sites of enzyme molecules or inside the membrane proteins, because of large effects of desolvation of buried charged groups and stronger interactions inside the protein interior. All of the above has motivated the development of algorithms that couple the treatment of conformational flexibility with the exchange of protons between the acidic and basic groups and the solvent. Examples include mean-field models with side-chain flexibility (You and Bashford 1995; Spassov and Bashford 1999; Barth et al. 2007), Monte Carlo sampling (Beroza and Case 1996; Georgescu et al. 2002), and advanced protocols for molecular-dynamics simulations (for review, see Mongan and Case 2005). Recently an implementation of replica-exchange method in constant pH dynamics (Khandogin and Brooks III 2006) demonstrated an improvement of the accuracy of this class of models that moves the predictions from first principles to a quantitative level. At the same time semi-empirical methods remain popular (Li et al. 2005), filling the existing gap between computational efficiency and the accuracy of compute-intensive physical approaches.

The main goal of the present work was to develop a fast and efficient computational tool to calculate a general set of protein properties related to the protonation of acidic and basic amino acid residues that can be used in diverse applications of protein modeling. Our intention was to replace FDPB with a Generalized Born (GB) model (Still et al. 1990) and to combine it with an iterative mobile clustering (IMC) approach to the binding of protons to molecules with multiple sites of titration (Spassov and Bashford 1999). After implementing a novel algorithm for preliminary optimization of hydrogen coordinates and a revision of the conception of model compounds, even in the simplest variant, without an explicit treatment of the structural relaxation, the computational method achieved high accuracy of pK predictions that was comparable to, or even better than, state-of-the-art methods reported in the literature. The combination of the GB and IMC approach (GB/IMC) provides a new computational protocol for pK calculations and modeling of the complex effects of protein ionization. The program we describe in this work includes modules to calculate the fractional protonation of titration sites and integral titration curves, the electrostatic contribution to the protein free energy, as well as the approximate folding energy as a function of pH. An additional module provides the option to reconstruct and refine the coordinates of all hydrogen atoms at a given pH. The optimization of the hydrogen-bond network includes assignment of hydrogen atoms according to calculated ionization characteristics, as well as a search for optimal tautomeric forms of histidine residues and the flipping of amide groups of Asn, Gln, and protonated carboxyl groups of Asp and Glu residues.

The method has been implemented as the ''Protein Ionization and pK prediction'' computational protocol in Accelrys Discovery Studio 2.0. In this work we provide validation results on a set of 24 proteins as well as several examples of complex modeling of pH-dependent properties of protein molecules, including membrane proteins.

Theory

The underlying physical formalism of our method is close to the model of protein ionization proposed by Bashford and Karplus (1990, 1991). It combines statistical physics with continuum electrostatics. A similar approach is used in many FDPB models of protein ionization (Bashford and Gerwert 1992; Yang et al. 1993; Antosiewicz et al. 1994; Schaefer et al. 1997; Gunner and Alexov 2000; Nielsen and Vriend 2001; Warwicker 2004) including membrane proteins (Karshikoff et al. 1994; Engels et al. 1995). However, the method proposed here has several novel features and algorithms that contribute to improvement of the accuracy of its predictions. Recently Onufriev et al. (2000) demonstrated that a modification of GB approximation can effectively replace the finite-difference technique to solve the Poisson–Boltzmann equation. We have implemented a similar variant of the GB solvation model (Dominy and Brooks III 1999a) extended to membrane proteins (Spassov et al. 2002) using CHARMm (Brooks et al. 1983). The GB model has been preferred not only because of computational efficiency but, equally important, to avoid the limitations of molecular size and shape specific for grid methods such as FDPB. In addition, the replacement of FDPB with a GB model allows us to substitute monopeptide model compounds with a library of precalculated polypeptide structures. This enables us to take advantage of recently reported accurate pK_a values estimated from potentiometric titration of natural acidic

and basic residues in blocked alanine pentapeptides (Thurlkill et al. 2006) and to use them as a consistent set of model pK's. In the work of Nielsen and Vriend (2001) and Georgescu at al. (2002) it has been demonstrated that an optimization of the hydrogen-bond network in the input X-ray structures can improve the accuracy of pK predictions of FDPB-based calculations. In our method a novel protocol for preliminary optimization of hydrogen positions has also been implemented. Our algorithm adds a search for optimal proton binding centers to a CHARMm protocol to construct and optimize the hydrogen coordinates.

Basics

A detailed recent description of the continuum electrostatics approach to protein ionization can be found in Bashford (2004). Here we will repeat briefly some of the key points, to avoid any confusion about the exact equations used in our implementation of the method.

A protein is modeled as a molecule with N titration sites that can exchange protons with the solvent. The ionization state l of a system is described by a microstate vector $\mathbf{X}_l$. With each component of $\mathbf{X}_l$, x_i, can take either the value 1 or 0 depending on whether site i is protonated or deprotonated. If the molecule is modeled as a single conformer, i.e., neglecting the coupling between the changes in the ionization state and the changes of protein structure, the probability to find the molecule in a microstate $\mathbf{X}_l$ at given pH is

$$\rho(\mathbf{X}_l, pH) = \frac{\exp[-G(\mathbf{X}_l, pH)/RT]}{\sum_{l=1}^{N} \exp[-G(\mathbf{X}_l, pH)/RT]} \quad (1)$$

where $G(\boldsymbol{X}_l, pH)$ is the free energy of the microstate $\mathbf{X}_l$. Note that the summation is carried out over all possible $N_p = 2^N$ microstates, which, even in a single conformer case, creates a serious combinatorial problem if the molecules had more than 15–20 titratable sites.

Almost all pH-dependent characteristics of a protein molecule, such as the titration of the individual ionogenic groups, the integral titration curves, or the pH-dependent contribution to the protein free energy, can be derived from the fractional protonation $\theta_i(pH)$ of individual sites of titration i:

$$\theta_i(pH) = \sum_{l=1}^{N_p} x_{i,l}\rho(\mathbf{X}_l, pH) \quad (2)$$

In the analysis below we will compare the experimental values of apparent pK_a, of individual residues with computed pK_{half} values. pK_{half} is defined as the pH value at which the fractional protonation of a titration site is equal to 0.5. However, in cases of abnormal shape of titration curves, pK_{half} must be used with caution and the pH dependence of fractional protonation should be regarded as the true ionization characteristic of protonation sites. It is convenient for the calculations if the microstate energy in Equation 1 is referenced to the completely deprotonated state (Bashford and Karplus 1991; Schaefer et al. 1997):

$$G(\mathbf{X}, pH) = \sum_{i=1}^{N_p} x_i \left[2.303RT(pH - pK_{\text{int }r,i}) + \sum_{j}^{N_p} W_{ij}q_j\right] + \sum_{i,j>i}^{N_p} x_i x_j W_{ij} \quad (3)$$

where $pK_{intr,i}$ are the calculated intrinsic pK's of the titratable groups assuming that all other residues in the protein are in their neutral state, according to the definition proposed by Tanford and Roxby (1972). In models with distributed formal charges over the multiple atoms of titratable groups, the interaction terms W_{ij} represent the additional energy to protonate site i if site j is protonated (Bashford and Gerwert 1992); q_i are the corresponding formal charges equal to -1 for acidic and +1 for basic groups. In the case of flexible molecules Equations 1–3 can be extended by adding pH-independent terms, which depend on conformational state (Spassov and Bashford 1999).

Calculating the electrostatic energy terms

All interaction terms in Equation 3 are calculated using the GB continuum electrostatic model. The use of GB models rather than grid-based techniques to solve the Poisson–Boltzmann equation not only significantly reduces computation time, but also makes the computations less dependent on molecular shape and size. The two main approximations in pure electrostatic approaches to proton binding are first, the assumption that all chemical contributions to the binding energy cancel each other in the thermodynamic cycle used in the definition of intrinsic pK's (see Equation 5 below) and second, that all nonbonded interactions of nonelectrostatic nature are insignificant compared to electrostatic interactions. The latter is not strictly true for short-range, H-bond type interactions. However, it is still a reasonable assumption considering that, in molecular mechanics force fields, such as in CHARMm, the hydrogen-bond interaction energy is often modeled by the sum of a van der Waals and a Coulombic term only and that, at nonoverlapping distances, the van der Waals contribution is negligible compared to the electrostatic interaction term. Similar to most of continuum electrostatic models, electrostatic energy terms in GB models are calculated as a sum of Coulomb term and a Generalized Born solvation energy term (Bashford and Case 2000). For this study,

we used the same variant of GB model as in Onufriev et al. (2000):

$$G_{el} = 332 \sum_i \sum_{j>i} \frac{q_i q_j}{\varepsilon_m r_{i,j}} - 166\left(\frac{1}{\varepsilon_m} - \frac{\exp(-K(I)F_{ij}^{GB})}{\varepsilon_{slv}}\right)\sum\sum\frac{q_i q_j}{F_{ij}^{GB}};$$

$$F_{ij}^{GB} = \sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2/4\alpha_i\alpha_j)} \quad (4)$$

$$K(I) = 0.316\kappa\sqrt{I}$$

where ε_m and ε_{slv} are the intramolecular dielectric constant and the dielectric constant of the solvent, respectively, and the α_i are effective Born radii. In difference to many other GB implementations, Equation 4 is extended with a correction for the ionic strength (Hawkins at al. 1995); an unchanged value of $\kappa = 0.7$ is used in all calculations.

Introducing libraries of model compounds

In most FDPB electrostatic models (Bashford and Gerwert 1992; Yang et al. 1993; Antosiewicz et al. 1994; Schaefer et al. 1997) the $pK_{int\ r}$ values are derived from calculations on model compounds with known experimental pK data according to a thermodynamic cycle described in Bashford and Karplus (1990):

$$pK_{int\ r} = pK_{mod} + (2.303RT)^{-1} \times [\Delta\Delta G(PH, P) - \Delta\Delta G(MH, M)] \quad (5)$$

where $\Delta\Delta G$ terms represent the noncovalent energy differences between the protonated (*PH* and *MH*) and deprotonated (*P* and *M*) forms of titratable atomic groups situated in the protein and model compound, respectively. Equation 5 is derived from a thermodynamic cycle applied to an ''alchemic'' process of transferring titratable groups to different structures and environments. The atomic groups of protonation sites are bonded to structures that are similar enough to cancel the ''chemical'' contributions to binding energy. However, the calculations of energy terms in Equation 5 depend not only on the chemical structure but also on the conformation of model compounds. The structures of model compounds is one of the key differences between our approach and other electrostatic models, where usually the model compound for each particular residue is represented by a monopetide compound with exactly the same atomic coordinates, as of the corresponding residue in the protein structure. In FDPB methods this is a necessary approximation to cancel the effects of numerical singularities due to the grid distribution of atomic charges. However, in the dense space of protein interior, it is not unusual for a titratable residue to obtain a conformation with intraresidue atomic contacts that are quite different from the contacts it would have in a relaxed structure in a water environment. Obviously, this difference is a potential source of error in predicted pK values and it introduces an inconsistency because of use of different model compound structures for residues of the same kind. One possible way to avoid such an inconsistency is to use reduced, nonpeptide reference compounds, for example, acetic acid instead of Glu or Asp monopeptides as in the MCCE method (Georgescu et al. 2002). However, in preliminary tests of the GB approach, our attempts to use the structure of acetic acid failed to reproduce the low pK_{half} values of carboxyl group in nonblocked Asp and Glu monopeptides. In a GB-based method it is not needed to use the same conformation of titratable residues in the protein and in the model compound. Based on this, we changed the concept of model compounds: In our method, we calculate the $\Delta\Delta G(MH, M)$ term in Equation 5 for a single, relaxed peptide structure, assumed as representative conformation of the model compound in solvent. In general, the monopeptide models used to calculate reference energies in Equation 5 can be replaced with any structure fulfilling the requirement of chemical similarity, for example, short polypeptides with known titration data. For the present study we took advantage of recently reported accurate NMR data for all natural acidic and basic amino acid residues in blocked pentapeptides of Ala-Ala-X-Ala-Ala type (Thurlkill et al. 2006). For each supported force field we constructed a library of low energy pentapeptide structures in β-strand conformation with the side chain optimized by sampling the side-chain dihedral angles and minimizing the CHARMm energy. The pK_{mod} values of pentapeptide model compounds are taken the same as determined by Thurlkill et al. (2006): 3.67 for X = Asp, 3.67 (a-carboxyl), 4.25 (Glu), 6.54 (His), 8.00 (a-amino), 8.55 (Cys), 9.84 (Tyr), and 10.40 (Lys), and a value of 12.0 is used for arginine.

A more realistic approach would be to construct representative ensembles of model compound conformations to derive the model compound free energy terms. However, during validation tests we achieved sufficient accuracy with the single conformation models and therefore we assumed them an appropriate approximation at the current stage of development.

Electrostatic free energy and pH stability

If the fractional protonation of the titration sites is known, an integration over the binding isotherms (Schellman 1975; Yang and Honig 1994) makes it possible to calculate the pH dependence of electrostatic free energy differences between any two conformers in a convenient

way. This approach can be used not only for comparison of different conformational states, but also for comparing different binding states, such as models with bound or unbound ligand. It can also be used in modeling of ensembles of conformers, weighted by the calculated absolute free energy (Spassov and Bashford 1999). For this purpose, the computational method reports the pH-dependent electrostatic contribution to absolute free energy calculated according to a convenient definition (Schaefer et al. 1997):

$$G(pH) = G(\infty) - 2.303RT \int_{pH}^{\infty} Q(pH)dpH \qquad (6)$$

where $G(\infty)$ is the electrostatic energy of the completely deprotonated state. The deprotonated state is constructed by removing the release hydrogen from all titratable groups. $Q(pH)$ can be either the total charge or the average number of bound protons calculated from the fractional protonations of individual sites. In the calculations the infinite pH value at the upper limit in the integral in Equation 6 is substituted with a value, pH_{inf}, estimated to be large enough to exclude the possibility of any numerically significant titration beyond that point. pH_{inf} is derived from the maximum possible contributions of charged groups to the pK shifts.

To provide a rough, but instant estimation of pH stability, the protocol reports the pH-dependent relative folding energy. The reference state used is the simplest, ''null'' model of unfolded state (Yang et al. 1993; Antosiewicz et al. 1994; Schaefer et al. 1997), assuming all groups in the unfolded state to titrate according to pK_a values adopted from the pentapeptide model compounds (Thurlkill et al. 2006) described above. Interestingly, the ''pentapeptide'' pK values of the acidic residues Asp ($pK_a = 3.67$) and Glu ($pK_a = 4.25$) are very close to the optimal pK_a values (Tan et al. 1995) estimated for a protein in the denatured state (3.6 and 4.0 for Asp and Glu, respectively). It implies that the pentapeptide pK_a values are better approximations to the unfolded state than the values of the model compounds of 3.9 and 4.3 used in Tan et al. (1995) and other papers.

IMC approach to protein ionization

It has been shown that the use of the iterative mobile clustering as in the IMC approach (Spassov and Bashford 1999) can be an effective way to treat the combinatorial problem arising from the exponential growth of protonation states with the number of titratable groups, not only in the single conformer cases, but also in systems with combined local and global conformational flexibility. The comparison with exact Boltzmann statistic calculations have also shown that IMC is not only much faster, but also more reliable on tested cases than a Monte Carlo sampling (Spassov and Bashford 1999). Therefore, we implemented the IMC algorithm to calculate the titration of acidic and basic groups. The computational program is capable of modeling not only rigid structures but also ensembles of distinct conformations in a way shown before (Spassov and Bashford 1999; Spassov et al. 2001). However, in the current study we report results using only the fast, single-conformer mode of the method.

Computational protocol

The equilibrium of proton binding can be very sensitive to small changes in the structure and model parameters. To minimize noise from possible variations of molecular models, we carried out all calculations with all heavy atoms fixed to the coordinates of the experimental structures and with standard force field values for all parameters, including the atomic radii and charges, with the exception of partial charges of neutral histidine, as explained below. The only physical parameter subject to parameterization in the present work is the internal dielectric constant ε_m as described in the Results and Discussion section. In principle, our approach allows the use of any CHARMm force field. At the current stage, we have validated the method for the CHARMm and CHARMm polar hydrogen force fields (Momany and Rone 1992).

The present method includes four main steps implemented as separate software components.

Step 1: Preliminary optimization and calculation of the effective Born radii

The first step is implemented as a program module written in CHARMm scripting language and includes a reconstruction and preliminary refinement of hydrogen atom positions followed by calculation of the atomic Born radii. The novel feature is the inclusion of a procedure for determination of the optimal proton binding centers on the carboxyl groups of Asp and Glu residues and C terminus.

The coordinates of hydrogen atoms are not present in most PDB files and, even if present, it is difficult to judge whether the data are objective, rather than the result of modeling. To avoid any possible dependency on the uncertainty in experimental hydrogen positions, initially all existing hydrogen atoms are stripped from the structure. Hydrogens are then reconstructed using the CHARMm HBUILD routine (Brünger and Karplus 1988) and energy optimized at fixed positions of all heavy atoms. In all steps of the algorithm the hydrogen building and optimization are carried out using the standard energy function of the selected force field. The next step invokes a search for optimal proton binding

center between the two oxygen atoms of titratable groups of any Asp and Glu residue or C terminus. For this purpose the potential at each atom of interest is calculated and the proton binding center is assigned to the atom embedded in the more negative electrostatic potential. In the following step, all missing hydrogens are constructed for all sites of titration, using the HBUILD routine. Finally, the GBIM CHARMm module (Spassov et al. 2002) is invoked to calculate the effective Born radii. GBIM is an extension of GBorn CHARMm module (Dominy and Brooks III 1999a) including an implicit membrane in the GB solvation model.

To avoid complications due to histidine tautomerism, we considered a simplified model of the neutral deprotonated state of the His side chain, assuming equal partial charges for both proton-binding nitrogens. The inspection of results from validation tests showed that in most cases the approximation of ''smeared charge'' works reasonably well. However, we believe there is room for improvement, especially to better reproduce cases with very large experimental pK shifts. This work is in progress.

Step 2: Calculating intrinsic pK and interaction energy terms

The Born radii calculated in step 1, along with the atomic coordinates of the molecule and model compounds, are used to evaluate the GB terms in Equation 4. This step is encoded as a separate program, GBPK, which computes all energy terms necessary to evaluate the intrinsic $pK_{int\ r}$ values and interaction terms included in the microstate energy described by Equation 3. The components of the interaction matrix W_{ij} in Equation 3 are calculated as described by Bashford and Gerwert (1992). In the calculations of $pK_{int\ r}$, GBPK uses coordinates of model compounds taken from structural libraries, as already described. The libraries are pre-generated separately for the CHARMm and CHARMm polar hydrogens (Momany and Rone 1992) force fields. The input data to GBPK include also the values of the intramolecular dielectric constant and the ionic strength of the solvent.

Step 3: Calculating the protein ionization

This step includes calculations of the fractional protonation of all sites of titration according to Equations 1–3. It also includes the evaluation of the total charge and the electrostatic contribution to the absolute free energy according to Equation 5 as well as the approximate estimation of the relative folding energy with respect to the ''null'' model of unfolded state as described above. The calculations are carried out by a program, PHPK, based on the IMC approach.

Step 4: Optimization of hydrogen-bond network

In addition to program modules to calculate the protein ionization, we developed a computational module for the assignment of hydrogen coordinates at a given pH and structural optimization of the entire hydrogen-bond network. Similar to step 1, the optimization algorithm is written as a CHARMm script.

The pK_{half} values computed in previous steps are used to define the protonation state of the titratable groups. If the selected pH value is less than the pK_{half} for an acidic or basic group, i.e., if the fractional protonation is >0.5, the structure of the group is generated in the protonated state and vice versa. The coordinates of all missing hydrogens are constructed by the CHARMm HBUILD routine. The next step includes a cycle over all sites of titration to optimize the geometry of groups with ambiguous binding center of H atom, such as OD1 or OD2 atoms of aspartate carboxyl groups, OE1 or OE2 of glutamate, C-terminal carboxyl groups or tautomeric forms of deprotonated histidines. The cycle also includes optimization of the ambiguous conformation of side-chain amide groups of Asn and Gln residues. The determination of the histidine tautomeric form is based on the comparison of the computed electrostatic potential at ND1 and NE2 atoms. The proton is assigned to the site with the larger negative potential. Two conformations of each protonated Asp, Glu, and C-terminal carboxyl groups, as well as each Asn and Gln are tested based on 180° rotations of side-chain carboxyl or amide groups, while the rest of the protein atoms are fixed and the low energy conformation is chosen. Note that the described flipping is the only case in all of the computations, which include manipulation of some heavy atoms. At the end of the cycle, the coordinates of all hydrogen atoms are re-optimized by energy minimization.

It has been shown recently that the interactions of side chains with backbone atoms are on average stronger than side-chain-to-side-chain interactions, with a very strong trend for polar Asp, Glu, Asn, Gln, Ser, and Thr residues (Spassov et al. 2007). Assuming the hypothesis of a dominant role of backbone atoms in determining the orientation of side chains of polar residues, we constructed our algorithm without sampling of mutual orientations of protein side chains avoiding a possibly substantial increase in computation time. We validated this approach in tests on several neutron-diffraction structures as discussed in the Results and Discussion section.

Results and Discussion

Lysozyme ionization and parameterization of the method

As discussed above, the only adjustable parameter in this study is the value of the molecular dielectric constant ε_m in Equation 4. The preliminary parameterization of ε_m has been carried out on a single structure—the structure of hen-egg white lysozyme (HEWL, PDB code 2lzt), a

structure widely used to test the predictive capabilities of computational approaches to protein ionization (Tanford and Roxby 1972; Imoto 1987; Spassov et al. 1989; Bashford and Karplus 1990; Nielsen and McCammon 2003; Mongan et al. 2004; and others). We mainly selected hen-egg lysozyme because of the complete list of existing experimental titration data for all residues of this molecule, excluding arginines. Also, the prediction of lysozyme ionization is challenging because of a number of residues with unusually high or low experimental pK_a values.

Figure 1 shows the accuracy of pK calculations as a function of ε_m. The model achieves the highest prediction accuracy between $\varepsilon_m = 10$ and 11 with an RMSD value <0.5 compared to recent results reported in the literature (Table 3, see below). The sharp decrease in accuracy at low ε_m is in agreement with many authors suggesting the use of high values of average dielectric permittivity when the flexibility of permanent dipoles and the mobility of charged groups are not modeled explicitly (Antosiewicz et al. 1994; Pitera et al. 2001; Simonson 2001). Interestingly, the optimal ε_m value, according to our GB model, is half the value $\varepsilon_m = 20$ obtained from FDPB calculations (Antosiewicz et al. 1994).

The preliminary tests on other structures showed a similar dependence on ε_m, and no systematic difference in accuracy has been observed between using $\varepsilon_m = 10$ or 11. We therefore used $\varepsilon_m = 11$ in all calculations in this study, with the exclusion of membrane proteins.

Table 1 shows the comparison between the experimental pK_a values and the calculated pK_{half} values for all titratable groups in a pH interval from 2 to 12. While both force fields, CHARMm and CHARMm polar H, lead to a correct estimation of the direction of the pK shifts for almost all residues, the all-hydrogen CHARMm models more accurately reproduced the unusually high experimental pK of Glu35 from the active site of the enzyme ($pK_a = 6.2$) (Bartik et al. 1994). On the other hand, the data show a better overall RMSD value for the CHARMm polar H than the CHARMm model. Note that most continuum electrostatic models, if successful in prediction of pK_a of buried groups such as Glu35, usually fail to predict the titration of surface groups, or vice versa (Bashford 2004). However, the difference between the two force fields is not drastic and in both cases the direction of pK shifts is predicted correctly for most of the residues. We were unable to find any reported results with better overall RMSD value than the RMSD of the calculations with CHARMm polar hydrogen force field in Table 1.

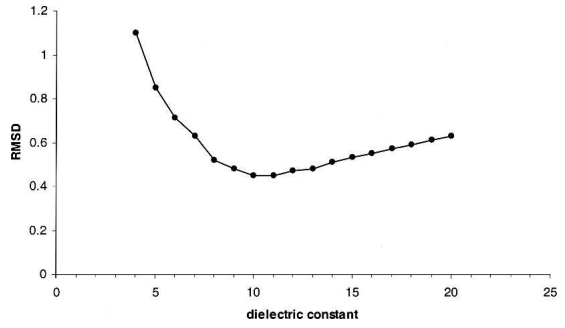

Figure 1. Average RMSD error of the predicted pK values of the acidic and basic residues as function of the value of internal dielectric constant. Experimental data from Kuramitsu and Hamaguchi (1980) and Bartik et al. (1994). The data are obtained on the example of the 2lzt PDB structure of HEWL.

The tests show that, without the initial optimization of hydrogen positions and using the standard set of pK_{mod}, instead of pentapeptide values, the calculated RMSD error in 2lzt case increases from 0.45 to 0.85. The increased accuracy of 0.27 pK units are from hydrogen optimization and the additional 0.13 pK units are due to the use of pentapeptide pK_{mod} values.

In addition to pK predictions we used the same lysozyme structure to test the ability of the method to predict some integral titration characteristics such as pH dependence of total charge and isoelectric point. Figure 2 shows the comparison between the calculated titration curves and the results of potentiometric titration at two different ionic strengths (Tanford and Roxby 1972). It is seen that at I = 0.1 M the calculated ionization curve is very close to the experimental data with some disagreement at low pH values. Although at I = 1.0 M the ''low pH'' disagreement increases, it is important that the calculated and the experimental curves show approximately the same pH dependence of the differences between the total charges at low and high ionic strength, suggesting that including the ionic strength dependence in Equation 4 is an appropriate approach. The calculations also give a perfect estimation of the lysozyme isoelectric point, pI = 10.21 versus the experimental value of 10.2 (Tanford and Roxby 1972).

Comparison of calculated pK_{half} values with experimental data for different proteins

To investigate the reliability of the method we tested it on a relatively large set of different protein structures. The set contains 24 crystallographic and NMR structures and a total number of 331 titratable groups with known experimental pK_a values. For each PDB entry in Table 2, the RMSD between calculated and experimental pK's of titratable groups is shown. Note that the calculations with both CHARMm force fields systematically show a high accuracy of the predictions with an RMSD of ~0.5 pK units, and only for a small number of outliners (1trs, 1rgg, and 1xnb) the RMSD is elevated, but still below one pK unit. In a recent comparative study by Davies et al. (2006) an empirical method, PROPKA (Li et al. 2005), appears to

Table 1. *Comparison of the computed pK_{half} values with the experimental pK_a values of the acidic and basic groups of HEWL*

	pK_{half}		
Residue	Experimental	CHARMm polar H	CHARMm
NTR1	7.9	7.81	8.00
LYS1	10.6	10.01	10.01
GLU7	2.9	3.17	3.39
LYS13	10.3	10.49	10.56
HIS15	5.4	6.20	5.87
ASP18	2.7	2.87	3.11
TYR20	10.3	10.85	11.18
TYR23	9.8	10.16	10.87
LYS33	10.4	10.58	10.79
GLU35	6.2	5.05	5.80
ASP48	2.5	2.96	2.91
ASP52	3.7	4.32	4.67
TYR53	>12	11.71	>12
ASP66	<2.0	2.15	2.87
ASP87	2.1	2.43	2.97
LYS96	10.7	11.18	11.42
LYS97	10.1	10.79	10.85
ASP101	4.1	3.89	3.92
LYS116	10.2	10.12	10.09
ASP119	3.2	3.08	3.28
CTR129	2.8	2.73	2.83
RMSD		0.45	0.57

The calculations were carried out on the 2lzt PDB structure using two different CHARMm (Momany and Rone 1992) force fields. Experimental data from Kuramitsu and Hamaguchi (1980) and Bartik et al. (1994).

be more accurate than some of the well known programs based on a continuum electrostatic model—MEAD (Bashford 1997), UHBD (Madura et al. 1995), and MCCE (Alexov and Gunner 1997). Contrary to the findings of the authors implying superiority of the empirical approach, the results in Tables 2 and 3 demonstrate that it is possible to achieve a considerably better accuracy, even better than those reported by the authors of PROPKA (RMSD 0.8–0.9), using a continuum electrostatic model of the same class as cited by the FDPB programs. Of equal importance is the fact that the use of the GB continuum model significantly reduces computational cost, compared to FDPB methods. According to the results of performance tests (Fig. 2) the CPU time used for a middle-size protein usually is about one minute, which is comparable to timings of the fastest empirical methods. Figure 3 also shows that the CPU timescales in an almost linear fashion with the size of proteins. The efficient scale-up is mostly because of the great reduction of sampling in the IMC algorithm, where only the number of mobile clusters increases with protein size, but the number of sites inside the clusters depending on an energy cutoff parameter stays small and not affected by protein size. In all studied cases a cutoff value of 0.5 kcal/mol was enough to achieve the prediction accuracy and to keep the maximum cluster size below 10–15 sites.

A smaller set of seven protein structures with well-characterized dissociation characteristics is widely used for test purposes by many authors. We used the same structures to compare the RMS error of GB/IMC calculations to several of the most accurate methods based on either physical or empirical approaches (Table 3). Note that GB/IMC not only shows better accuracy, as seen in Table 3, but also does that at very low computational cost, if compared to more sophisticated physical methods such as constant pH dynamics or MCCE. The accuracy is even better than the accuracy of the empirical SCP approach, in which results are achieved after intensive fitting of multiple parameters.

Modeling the pH stability

Some examples of the computed pH-dependent contribution to the folding energy comparing to the data from unfolding experiments in urea are given in Figure 4. It is interesting to note that, although the pH dependence of relative folding energy is calculated using the simplest ''null'' model of unfolded state (Schaefer et al. 1997), i.e., ignoring the possible effects of any resting electrostatic interactions in unfolded state, in most cases the results show a reasonable estimation of the pH stability which is comparable to the results obtained from more realistic models of the electrostatic interactions in the unfolded state, such as the native-like structures modeled by Elcock (1999). In the Rnase T1 example shown in Figure 4, the method is unable to predict sufficiently well the shape of the denaturation curve but gives a close estimation of the pH optimum.

The implementation also allows the curves of pH stability to be obtained using other simple models, for

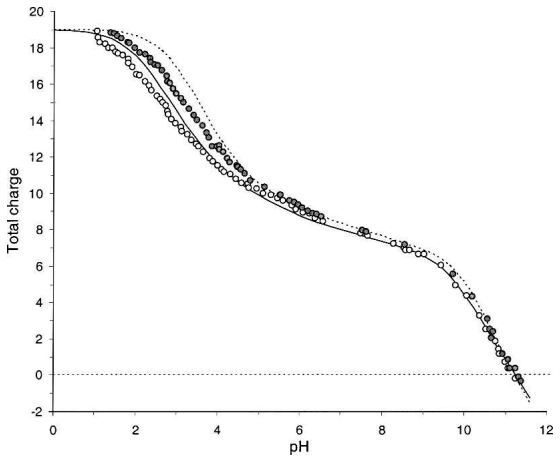

Figure 2. The pH dependence of lysozyme total charge at different ionic strengths. Open circles and solid line: experiment and computed values at 0.1 M ionic strength; filled circles and broken line: the corresponding curves at 1.0 M ionic strength.

Table 2. *The RMSD between calculated and experimental pK_a values obtained on a test set of 24 protein structures*

PDB code	Protein	Ng	CHARMm polar hydrogens	CHARMm all hydrogens
4pti	Bovine pancreatic tripsin inhibitor	14	0.36 (0.9)	0.36 (0.8)
2lzt	Hen-egg lysozyme	21	0.45 (0.8)	0.57 (1.3)
2rn2	RNase H	25	0.59 (1.2)	0.68 (1.6)
3rn3	RNase A	16	0.44 (1.1)	0.63 (1.3)
1pga	B1 binding domain of protein G	15	0.42 (1.2)	0.57 (1.2)
3icb	CabD	10	0.33 (0.7)	0.35 (0.8)
1hng	Cell adhesion molecule CD2	14	0.55 (1.7)	0.53 (1.3)
1a2p	Barnase	12	0.60 (1.3)	0.49 (0.8)
1omu	Turkey ovomucoid third domain	15	0.64 (1.4)	0.70 (1.3)
9rnt	RNase T1	14	0.54 (1.2)	0.65 (1.5)
1bi6	Bromelain inhibitor IV heavy chain	18	0.54 (1.3)	0.53 (1.2)
1bi6	Bromelain inhibitor IV light chain	4	0.18 (0.3)	0.27 (0.5)
1rgg	RNase Sa	24	0.83 (2.7)	0.89 (1.7)
1igd	B2 domain of protein G	16	0.35 (0.5)	0.36 (0.9)
135l	Turkey lysozyme	11	0.63 (1.1)	0.65 (1.1)
1div	N-terminal domain of L9	6	0.26 (0.5)	0.32 (0.6)
1xnb	Xylanase	13	0.70 (1.2)	1.09 (1.2)
1kxi	Cardiotoxin A5	3	0.57 (1.0)	0.50 (0.8)
1beo	Criptogein	10	0.46 (1.0)	0.56 (1.3)
1trs	Thioredoxin (oxidized)	17	0.88 (3.1)	0.71 (1.0)
1qbs	HIV-1 protease (DMP-323 complex)	16	0.34 (0.7)	0.34 (0.7)
1de3	α-Sarcin	25	0.66 (1.3)	0.70 (1.3)
2bus	Bull seminal inhibitor	4	0.46 (0.9)	0.49 (0.9)
1egf	Epidermal growth factor	9	0.49 (0.9)	0.53 (0.8)
Average RMSD		331	0.51	0.56

Ng is the number of titratable groups with known experimental data. The maximum error is shown in brackets.

example calculating the difference between the absolute electrostatic energies of native structure and a relaxed structure in β-strand conformation (Schaefer et al. 1997). Here we do not provide such data because during the tests the null model systematically showed a better overall fit to experimental curves. However, at extreme pH, an extended conformation could be a reasonable model because the noncompensated repulsion between the groups of the same charge will be reduced.

Membrane proteins: bacteriorhodopsin ionization

The modeling of ionization characteristics of membrane proteins is important for the understanding of molecular mechanisms, and many of them are of high interest for pharmacology and bioenergetics, such as the regulation of ion channels, the activation of membrane receptors or utilization of light energy. Several studies demonstrate that membrane environment can be successfully approximated by including an implicit low dielectric slab in FDPB ionization models (Karshikoff et al. 1994; Engels et al. 1995) However, despite the availability of several GB solvation models with implicit membrane (Spassov et al. 2002; Im et al. 2003; Tanizaki and Feig 2006), none of the models, to our knowledge, have been used in pK calculations. To test the possibility of replacing grid-based FD calculations with the fast GB code, we used the GBIM CHARMm module to include an implicit membrane and to repeat some of the calculations reported in the MEAD-based study of ionization of bacteriorhodopsin (BR) (Spassov et al. 2001). For the sake of comparison we used the same protein model, based on the 1c3w X-ray structure of ground state. Here we compare the results from the simplest variant of the BR ground state, without including O_2H_5 ions or larger water clusters as possible sites of proton storage as in the more realistic model proposed by the authors. A more detailed modeling of bacteriorhodopsin ionization and the complex mechanism of proton transfer was beyond the goals of this study and can be found elsewhere (Engels et al. 1995; Sampogna and Honig 1996; Spassov et al. 2001; Song et al. 2003; Ferreira and Bashford 2006). A common problem for all "two dielectric" models with implicit membrane is the choice of the molecular dielectric constant, because of the requirement for this value to be the same in both membrane and protein interior. A low value of $\varepsilon_m = 2$–4 is suitable for a dielectric slab formed by lipid tails, but inconsistent with the optimal value of 10–11 obtained above on a set of 24 water-soluble proteins. Because a significant fraction of membrane proteins is exposed to the solvent, a compromised value is needed to represent the dielectric properties

Table 3. *Comparison of the accuracy of pK predictions with other methods*

	Ngr	GB/IMC	MCCE	Const. pH	FD/DH	SCP	PROPKA
4pti	14	0.36	0.47	NA	0.35	0.33	0.6
2lzt	21	0.45	0.76	0.6	0.47	0.49	0.66
2rn2	25	0.59	0.87	0.9	1.17	0.57	0.72
3rn3	16	0.44	0.66	1.2	0.87	0.55	0.94
1pga	15	0.42	0.63	NA	0.80	0.59	0.72
3icb	10	0.33	0.38	NA	0.37	0.39	0.9
3rnt	4	0.28	0.54	NA	NA	0.41	NA
Average		0.41	0.63	—	0.67	0.49	0.76

MCCE: Georgescu et al. (2002); Const. pH: Khandogin and Brooks III (2006); FD/DH: Warwicker (2004); SCP: Mehler and Guarnieri (1999); PROPKA: Li et al. (2005).

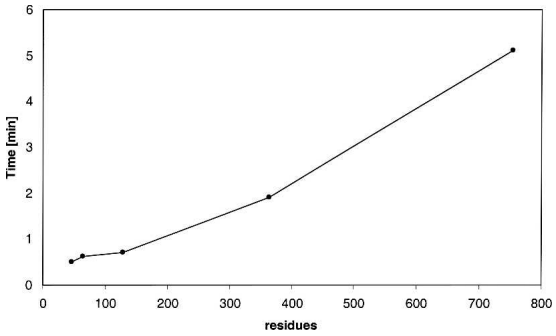

Figure 3. The CPU time used to calculate the ionization of proteins with different chain lengths. The data were generated on an Intel Pentium4 3.0 GHz machine.

of both transmembrane region and the molecular interior outside the membrane. We found that a value of $\varepsilon_m \sim 28$–10 reproduces the balance of electrostatic interactions in BR molecule well enough, based on the comparison of calculated pK_{half} of the residues in the transmembrane region with the experimental data shown in Table 4. It is seen that GB/IMC results are no less accurate than the results obtained in the previous FDPB calculation on the same model. Note that removing the membrane dramatically changes the proton affinities of some functionally important residues, such as Asp85 and Asp212, as well as the titration behavior of the proton release dyad Glu194–Glu204. We consider the results in Table 4 to be quite encouraging and believe this extension of the GB/IMC method to membrane proteins can be very helpful in modeling the membrane proteins, despite some issues regarding the uniform dielectric environment in the GB implicit solvent model.

HIV-1 protease

In an attempt to demonstrate the complete functionality of the method, the GB/IMC implementation was tested on HIV-1 protease using several different experimental structures. This enzyme was selected, not so much because of its biomedical importance, but because numerous titration experiments of different kinds are reported in the literature (Trylska et al. 1999). The calculations were carried out using three X-ray structures, one for apo-enzyme and two complexes with different inhibitors, DMP-232 and KNI-272 (Table 5). KNI-272 is modeled as an ionizable group—it titrates in water with a pK_a of 4.9 (Wang et al. 1996).

From Table 5, the data show very good agreement between experimental and calculated pK_{half} values for apo-enzyme and the complex with the symmetric DMP-232 inhibitor. The single, but important disagreement is the KNI-272 case where our calculations do not predict a low value of the first pK_{half} of catalytic dyad formed by Asp25 residues from A and B subunits as suggested by Wang et al. (1996). To study the problem further, we used several models, including models with conformational flexibility and incorporated water molecules, but we have been unable to find any structural or other reasons for the low pK of Asp 25.

In Figure 5 the stability curves computed by using two, ''null'' and ''beta,'' reference models of unfolded state (see Materials and Methods) are compared to the experimental data from unfolding in urea.

In contrast to the FDPB methods, using the GB models makes it technically possible to model the energy of binding between two molecular units in a very easy manner. This can be done simply by calculating the difference of the absolute electrostatic contribution to the free energy, $\Delta\Delta G_{bind}$, between (Equation 5) the native structure and a structure with the units separated on long distance, say 200 Å. Figure 6 shows the calculated pH dependence of negative energy of binding ($-\Delta\Delta G_{bind}$) of KNI-272 inhibitor to HIV-1 protease, compared to the corresponding energy curve derived from the experimental pH-dependent association constant. It is seen that the calculation reproduces remarkably well the pH optimum of binding as well as the association of the inhibitor in the

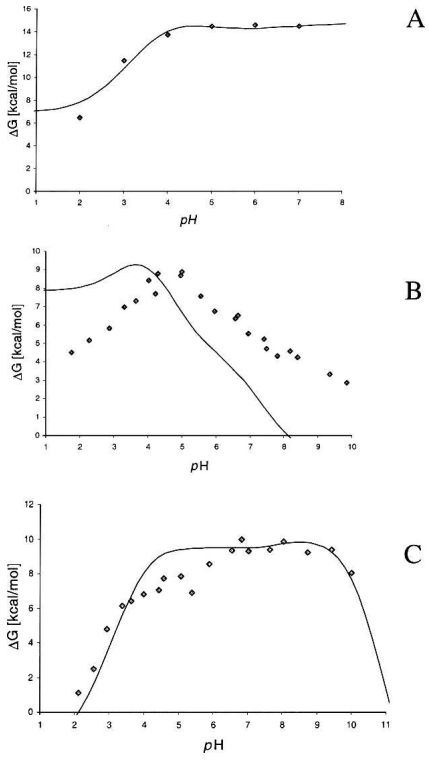

Figure 4. The comparison of calculated free energy of unfolding compared to the data from urea denaturation. (*A*) Hen-egg lysozyme; (*B*) Rnase T1; (*C*) Rnase A. Experimental data from Pfeil and Privalov (1976) and Pace et al. (1990).

Table 4. *Comparison of calculated and experimental pK_{half} values of the titratable residues from the Bacteriorhodopsin interior*

	pK_{half}			
	Calculated	Calculated without membrane	Calculated using MEAD with membrane	Experiment
ARG82	>14	>14	>14	>13.8
ASP85	3.3	8.4	1.7	2.6
ASP96	11.5	9.8	>14	>12
ASP115	7.2	9.6	8.4	>9.5
GLU194	>12	9.2	>14	Proton release
GLU204	3.3	7.6	<0	dyad; pK ~9.5
ASP212	<0.00	8.8	<0	<2.5
Schiff base	>14	>14	>14	>12

List of experimental data is the same as in Spassov et al. (2001).

low pH region. The observed disagreement at elevated pH could be due to the fact that possible changes in oligomeric state with pH are not taken into account.

Figure 7 shows the hydrogen bonds between residues from the active site and the inhibitor obtained after the final step of the computations on DMP-323 complex. It is notable that the predicted configuration of hydrogen bonds is exactly the same as suggested earlier from an analysis of the experimental data (Yamazaki et al. 1994).

Structural optimization of hydrogen-bond networks

Most of the protein structures in the PDB are solved without enough resolution to determine the position of the hydrogen atoms. Only a small number of neutron-diffraction and high-resolution X-ray structures contain hydrogen coordinates obtained from the diffraction maps. Because the positions of hydrogen atoms, especially polar hydrogens, can be of significant importance for many atomistic models, numerous algorithms have been developed to construct proton positions from the coordinates of heavy atoms. One of the most widely used tools for building and refinement of a hydrogen-bond network is the CHARMm HBUILD routine (Brünger and Karplus 1988). Recently, several new methods have addressed hydrogen-bond optimization by incorporating some more structural features such as the tautomeric forms of deprotonated histidines or the ambiguous orientation of Asn and Gln side-chain amide groups (Hooft et al. 1996; Word et al. 1999). Two recent methods of this class also include pH dependency in the optimization algorithm (Labute 2007; Li et al. 2007). Interestingly, both of the new methods use a search for a microstate with minimal energy for assignment of protons to titratable groups. In principle, the microstate approach can be a reasonable approximation when the contribution of a single state to the partition sum in Equation 1 is strongly dominant. However, in many interesting cases the use of optimal microstate energy carries the risk of producing misleading results, because in certain conditions, the molecular system can occupy the state with minimal energy with a low probability. Also, a microstate of protonation cannot be observed experimentally, while the calculated fractional protonations are the quantities that correspond directly to both experimental titration data and to the existence of the protons in neutron-diffraction structures of titratable residues. In the present work we used a completely different approach. Instead of the ''optimal'' microstate, the hydrogen assignment is based on the average fractional protonation of the residues obtained in the calculations of ionization equilibrium, as is described in the Computational protocol/step 4 section.

Table 6 summarizes the results of the refinement of the network of hydrogen interactions obtained on a set of five neutron-diffraction structures crystallized at different pH. In the comparison of the hydrogen structure of individual residues we adopted a criterion used recently for similar purposes (Labute 2007): A prediction is assumed to be correct if the ionization states are the same, the tautomeric form or the orientation of flipping groups is the

Table 5. *Computed pK_{half} of the acidic residues of HIV-1 protease compared with experimental data for apo-enzyme and two inhibitor complexes*

	pK_{half}		
		Computed	
		Subunit A	B
Residue	Experiment		
	Apo-enzyme, 1hhp structure		
Asp25	<5.9	5.3	5.4
	Complex with DMP-323 inhibitor, 1qbs structure		
Asp 25	>7.2	7.3	7.4
Asp 29	2.0	2.4	2.6
Asp 30	4.0	3.5	3.4
Asp 60	3.1	3.3	3.3
Glu 21	4.5	4.3	4.3
Glu 34	4.9	4.7	4.7
Glu 35	3.7	3.6	3.6
Glu 65	3.7	3.5	3.5
	Complex with KNI-272 inhibitor, 1hpx structure		
Asp 25	A B	8.7	6.3
Asp 29	>6.2 < 2.5	2.5	2.3
Asp 30	<2.5 < 2.5	3.5	3.7
Asp 60	3.9 3.8	2.9	2.5
	3.0 3.0		
KNI-272	2.9	3.9	

Experimental data from Yamazaki et al. (1994), Wang et al. (1996), Smith et al. (1997), Velazquez-Campoy et al. (2000).

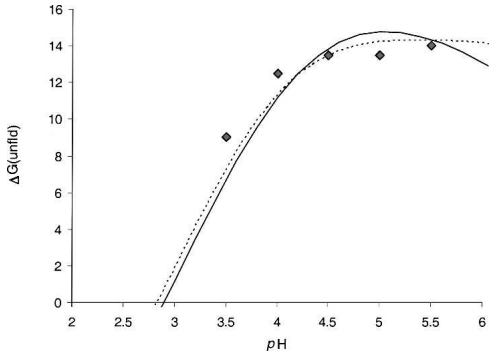

Figure 5. The pH-dependent contribution to unfolding energy (kcal/mol) of HIV-1 protease compared to the data from unfolding in urea (Todd et al. 1998). Solid line: data obtained using the null model; broken line: using β model of unfolded state.

same, and the hydrogen atoms are rotated by no more than 60°. In rare cases, some groups of hydrogen atoms are completely missing in crystallographic coordinate lists, such as all protons from the lysine NH3 group. In these cases we assume the prediction as correct, because it does not contradict with an experimental estimation.

While comparison to neutron-diffraction structures is the appropriate method to validate the positions of hydrogen atoms (Brünger and Karplus 1988), these structures have not been used in recent studies, and high resolution X-ray structures (Labute 2007) or quite indirect testing (Li et al. 2007) are used instead. In a previous study on the same subject (Hooft et al. 1996) the authors discard the neutron-diffraction structures as inappropriate for validation of the optimized hydrogen positions, suggesting experimental problems as a reason for too many cases of poor predictions. In contrast, our results demonstrate a good agreement between the predicted and neutron-diffraction position not only for hydrogen atom positions but also for the calculated and observed protonation states of titratable residues as can be seen in Tables 6 and 7. A closer inspection of structures shows an extremely good fit of nonpolar hydrogen atoms with very rare cases of differences >0.2–0.3 Å. Those differences are observed in the hydrogen positions of some methyl groups rotated occasionally by ~50°–60°, such as in Leu49 and Leu89 in 2l2k and Val196 in 2gve structures. Note, however, that the rotations of the symmetric CH3 or NH3 groups are usually insignificant from the modeling point of view. As could be expected, the positions of polar atoms show more differences, especially on the surface, but even in these cases the agreement to neutron-diffraction structures is quite good. The inspection of the 2gve structure of xylose isomerase, the largest of the studied proteins, shows that, out of 114 groups, the position of the polar protons are predicted correctly for 93 (82%), with differences coming mostly from the Ser and Thr residues situated on the surface of the protein.

In Table 7 the comparison between the calculated protonation states and the protonation states determined from the neutron-diffraction structures can be seen. The data provide two possible ways to make the comparison—from a strict prediction based on the calculated pK_{half} values, and from an analysis based on the calculated fractional protonation corresponding to the experimental pH value. An experimental state is defined as protonated if the corresponding hydrogen atom is present in the PDB coordinate list, and vice versa. For each entry only the nontrivial cases are listed, discarding the residues that are unlikely to titrate because of large differences between experimental pH and model pK's.

When using neutron-diffraction structures for comparison, it must be noted that, for a proton to be observed, it is not absolutely necessary that the occupancy be more than one-half. If we assume that states with reasonable fractional protonation, say no less than 0.25, can be seen on neutron-diffraction maps, then the predicted states of 31 from 36 residues, or 86%, can be assumed as consistent with neutron-diffraction structure. Even using the stricter criterion of 50% occupancy, a high percentage (75%) of the states is predicted correctly. It is notable that, in lysozyme, not only is the protonation state of Glu35 from active site predicted correctly, but the hydrogen, HE2, is bound to the same oxygen atom as the deuterium atom, DE2 in the experimental structures (Fig. 8), and has a similar orientation. The only difference observed in the 1lzn structure is in the protonation of Glu7. The observed protonated state in 1lzn structure

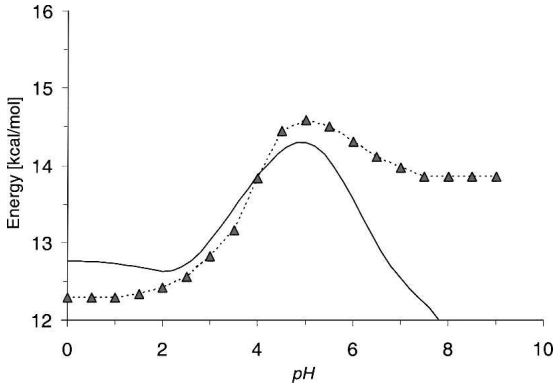

Figure 6. The pH-dependent contribution to the binding energy of KNI-272 inhibitor to HIV-1 protease, compared to the experimental values of association constant K_a taken from Velazquez-Campoy et al. (2000). The solid line represents the computed values of $-\Delta\Delta G_{bind}$. The triangles correspond to the values of $2.303RT log K_a$.

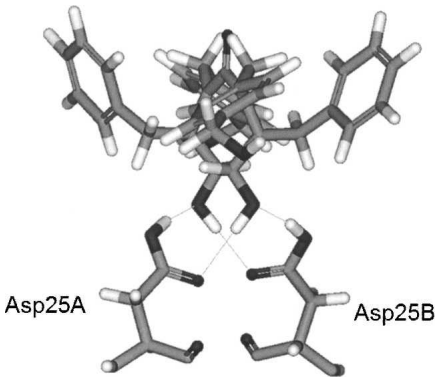

Figure 7. The predicted hydrogen bonds in the complex of HIV-1 protease with DMP-232 inhibitor.

does not agree with the predicted state, and also with the low pK_a of 2.9 observed in NMR experiments in water (see Table 2). It is tempting to speculate that the difference is related to the effect of the crystal field, but we will leave such a hypothesis for further investigation.

The most striking result of hydrogen optimization is the 100% prediction of the tautomeric states of all 10 deprotonated forms of histidine residues in the test set as shown in Table 7. It is not trivial to predict the only one from 2^{10} or 1024 possible combinations. We regard this result as valuable, because the right histidine tautomers can be critical for important implementations of method such as ligand docking. Figure 9 illustrates a prediction of histidine protonation on the example of myoglobine, 1l2k structure.

Conclusion

The results of several studies show that, without an explicit treatment of conformational flexibility, the traditional methods based on continuum electrostatic models have accuracy problems in the prediction of the dissociation characteristics of titratable residues in proteins (Simonson 2001; Bashford 2004; Khandogin and Brooks III 2006). The existing gap between the accuracy and computational efficiency of direct physical approaches motivates the development of empirical methods (Li et al. 2005) recently suggested as more accurate (Davies et al. 2006) than most well known FDPB programs. In contrast, the results of this study demonstrate that it is possible to create an accurate and computationally effective approach to protein ionization based entirely on the traditional continuum electrostatic model. The proposed GB/IMC approach systematically yields results that appear to be superior not only to the fast empirical methods, but also to the results of all other methods we found in literature, some of them based on much more detailed physical models. Several factors incrementally contributed to the improvement of the method, such as the replacement of monopeptide model compounds with structural libraries of polypeptides and the novel algorithm for preliminary optimization of the hydrogen structure. Among the other factors contributing to high accuracy, we like to mention the efficiency of the IMC approach, as well as the use of CHARMm GB models with improved accuracy (Dominy and Brooks III 1999a,b), and the well-balanced CHARMm (Momany and Rone 1992) force field atomic parameters. The implementation of GBIM (Spassov et al. 2002) CHARMm module in the calculations of the effective Born radii makes GB/IMC, to our knowledge, the first GB-based approach to protein ionization that is applicable to membrane proteins.

The proposed algorithm for structural optimization of hydrogen-bond networks also shows a reasonable agreement with the crystallographic protonation states and position of hydrogen atoms in neutron-diffraction structures. It is notable that the high percentage of correct predictions is achieved using input structures that are completely stripped of all hydrogen atoms. We believe this method will be valuable for future implementation in molecular dynamics protocols, ligand docking, and

Table 6. *Comparison between the predicted hydrogen position and the neutron-diffraction structures*

PDB code	Protein	pH	Nres	Ncorr	Flips		His	
					Total	Correct	Total	Correct
1vcx	Rubredoxin	8.0	53	53	1	1	NA	NA
1lzn	HEWL	4.7	129	119	18	14	NA	NA
1l2k	Met-myoglobin	6.8	151	146	5	4	6	6
2gve	Xylose isomerase							
		8.0	388	370	21	17	4	4
6rsa	Rnase A	6.6	128	116	17	15	NA	NA
	Total % predicted		849	802 94%	62	51 82%	10	10 100%

Ncorr, number of residues with correctly predicted hydrogen position; Flips, comparison of the orientations of amide or carboxyl groups of Asn, Gln, Asp, and Glu residues; His, comparison of tautomeric forms of deprotonated histidine residues.

Table 7. *Comparison between calculated and experimental protonation states in neutron-diffraction structures*

1lzn, pH 4.7			1l2k, pH 6.8			2gve, pH 8.0			6rsa, pH 6.6		
ASP18	3.66	**D**	NTR1	7.30	NA	NTR1	7.6	P[a]	NTR1	7.40	**P**
	0.13			0.75			0.30			0.86	
ASP48	2.80	**D**	HIS12	6.76	**D**	HIS49	6.17	**D**	HIS12	6.86	**P**
	0.03			0.48			0.02			0.62	
ASP52	4.54	**D**	HIS24	6.69	**D**	HIS54	7.6	P[a]	HIS48	8.70	**P**
	0.47			0.47			0.30			0.99	
ASP66	3.67	**D**	HIS36	7.19	**P**	HIS71	7.03	**D**	HIS105	6.95	**P**
	0.13			0.69			0.11			0.68	
ASP87	3.33	**D**	HIS48	6.22	P[b]	HIS96	5.13	**D**	HIS119	6.50	P[a]
	0.07			0.22			0.03			0.43	
ASP101	3.90	**D**	HIS64	4.47	**D**	HIS198	6.64	P[b]		1vcx, pH 8	
	0.18			0.02			0.06				
ASP119	3.45	**D**	HIS81	6.37	NA	HIS220	7.08	P[b]	NTR1	9.22	P
	0.08			0.31			0.15			0.94	
GLU7	3.70	P[b]	HIS82	6.41	**D**	HIS230	6.67	P[b]			
	0.13			0.33			0.06				
GLU35	5.67	**P**	HIS97	6.28	**D**	HIS243	6.40	**D**			
	0.89			0.26			0.07				
HIS15	7.50	**P**	HIS113	5.60	NA	HIS285	9.35	**P**			
	0.99			0.10			0.93				
CTR129	2.90	**D**	HIS116	6.71	NA	HIS382	7.54	P[a]			
	0.03			0.46			0.29				
			HIS119	4.94	**D**						
				0.19							

Numerical data: first row, computed pK_{half} values; second row, the fractional protonations of residues. P, residue protonated in crystal structure; D, deprotonated; NA, more than one polar hydrogen is missing. In boldface, accurately predicted structures.
[a] Underpredicted, but close.
[b] Completely incorrect prediction.

protein docking algorithms, as well as in hybrid protocols including molecular dynamics or quantum mechanics calculations.

The integration of the GB/IMC program modules within Discovery Studio provides a convenient graphical tool that saves a lot of effort usually needed in the preparation of input data. The Protein Ionization and pK prediction protocol is already in the hands of many investigators and the authors are hopeful that it will be useful in many areas of protein modeling.

Materials and Methods

Data sets

A set of 24 proteins has been constructed for the purpose of accuracy tests of pK predictions (Table 2). The entries include X-ray structures representing proteins with pK_a values of a relatively large number of residues determined by NMR experiments. The set contains the proteins from the collection of Forsyth et al. (2002) extended with some proteins taken from MCCE and PROPKA studies (Georgescu et al. 2002; Li et al. 2005). A small number of PDB entries (1hic, 1mhi, 1a91, and 2ci2) are discarded because of an incomplete list of atomic coordinates or too many residues with ambiguous titration. Most of references to the sources of experimental pK values can be found in the literature (Forsyth et al. 2002; Georgescu et al. 2002; Li et al. 2005). Additionally, the experimental pK list is updated with data for more residues, found after a search in pK_a database PPD (Toseland et al. 2006).

A second set of five neutron-diffraction structures has been selected for testing the optimization of the hydrogen position. The selection (see Table 6) includes representatives of all neutron-diffraction structures deposited in PDB with resolution

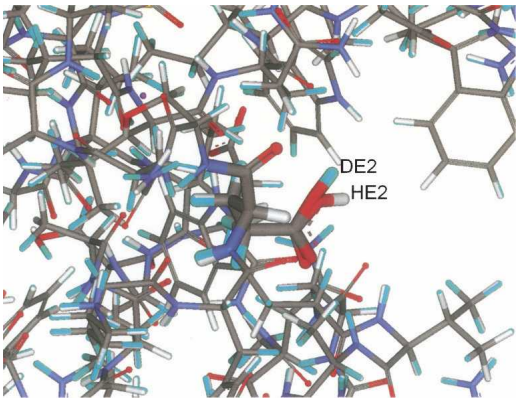

Figure 8. The differences between predicted and neutron-diffraction hydrogen positions in the active site of HEWL (1lzn). In boldface, the structure of Glu35. In light-blue are the hydrogen and deuterium atoms in neutron-diffraction structure.

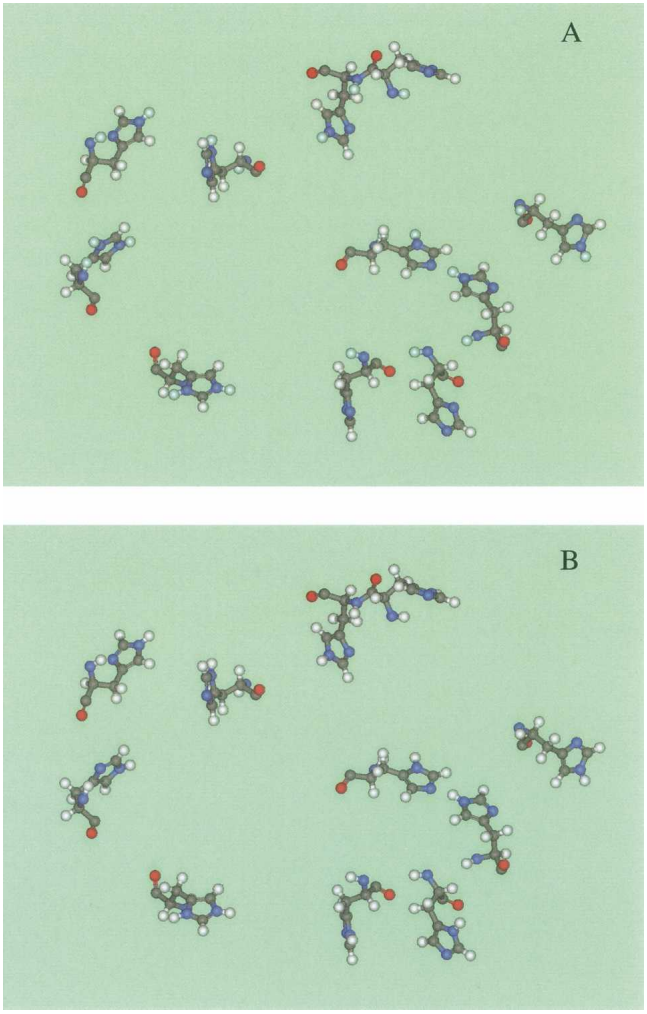

Figure 9. The protonation and tautomeric states at pH 6.8 of the histidine residues in 1l2k structure of met-myoglobin. (*A*) Predicted. (*B*) Neutron-diffraction structure.

<2.0 Å and with reported pH of the solvent during the crystallization.

Acknowledgments

We thank Drs. Paul Flook, Dipesh Risal, Marc Fasnacht, Yi-Shiou Chen, and Eric Yan for their help and useful discussions.

References

Alexov, E.G. and Gunner, M.R. 1997. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys. J.* **72:** 2075–2093.

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

Barth, P., Alber, T., and Harbury, P.B. 2007. Accurate, conformation-dependent predictions of solvent effects on protein ionization constants. *Proc. Natl. Acad. Sci.* **104:** 4898–4903.

Bartik, K., Redfield, C., and Dobson, C.M. 1994. Measurement of the individual pKa values of acidic residues of hen and turkey lysozymes by two-dimensional ^{1}H NMR. *Biophys. J.* **66:** 1180–1184.

Bashford, D. 1997 An object-oriented programming suite for electrostatic effects in biological molecules. In *Lecture notes in computer science*. (eds. Y. Ishikawa et al.) Vol. 1343, pp. 233–240. Springer, Berlin.

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

Bashford, D. and Case, D. 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51:** 129–152.

Bashford, D. and Gerwert, K. 1992. Electrostatic calculations of the pK values of ionizable groups in Bacteriorhodopsin. *J. Mol. Biol.* **224:** 473–486.

Bashford, D. and Karplus, M. 1990. pK_a's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochemistry* **29:** 10219–10225.

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

Beroza, P. and Case, D.A. 1996. Including side chain flexibility in continuum electrostatic calculations of protein titration. *J. Phys. Chem.* **100:** 20156–20163.

Brooks, B., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4:** 187–217.

Brünger, A.T. and Karplus, M. 1988. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins* **4:** 148–156.

Davies, M.N., Toseland, C.P., Moss, D.S., and Flower, D.R. 2006. Benchmarking pKa prediction. *BMC Biochem.* **7:** 18. doi: 10.1186/1471-2091-7-18.

Dominy, B.N. and Brooks III., C.L. 1999a. Development of a Generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **103:** 3765–3773.

Dominy, B.N. and Brooks III, C.L. 1999b. *Parameterization of a Generalized Born model for the MSI CHARMm Momany and Rone force field.* Accelrys Document, San Diego, CA.

Elcock, A.H. 1999. Realistic modeling of the denatured states of proteins allows accurate calculations of the pH dependence of protein stability. *J. Mol. Biol.* **294:** 1051–1062.

Engels, M., Gerwert, K., and Bashford, D. 1995. Computational studies of the early intermediates of the bacteriorhodopsin photocycle. *Biophys. Chem.* **56:** 95–104.

Ferreira, A.M. and Bashford, D. 2006. Model for proton transport coupled to protein conformational change: Application to proton pumping in the bacteriorhodopsin photocycle. *J. Am. Chem. Soc.* **128:** 16778–16790.

Fitch, C.A. and García-Moreno, E. B. 2006. Structure-based pKa calculations using continuum electrostatics methods. *Curr. Protoc. Bioinformatics* **8:** 8.11.

Forsyth, W.R., Jan, M., Antosiewicz, J.M., and Robertson, A.D. 2002. Empirical relationships between protein structure and carboxyl pKa values in proteins. *Proteins* **48:** 388–403.

Georgescu, R.E., Alexov, E.G., and Gunner, M.R. 2002. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophys. J.* **83:** 1731–1748.

Gunner, M.R. and Alexov, E. 2000. A pragmatic approach to structure based calculation of coupled proton and electron transfer in proteins. *Biochim. Biophys. Acta* **1458:** 63–87.

Hawkins, G.D., Cramer, C.J., and Truhlar, D.G. 1995. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **246:** 122–129.

Honig, B. and Nicholls, A. 1995. Classical electrostatics in biology and chemistry. *Science* **268:** 1144–1149.

Hooft, R.W., Sander, C., and Vriend, G. 1996. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. *Proteins* **26:** 363–376.

Im, W., Feig, M., and Brooks III, C.L. 2003. An implicit membrane Generalized Born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys. J.* **85:** 2900–2918.

Imoto, T. 1987. Electrostatic free energy of lysozyme. *Biophys. J.* **44:** 293–298.

Juffer, A.H. 1998. Theoretical calculations of acid dissociation constants of proteins. *Biochem. Cell Biol.* **76:** 198–209.

Karshikoff, A., Spassov, V., Cowan, S.W., Ladenstein, R., and Schirmer, T. 1994. Electrostatic properties of two porin channels from *Escherichia coli*. *J. Mol. Biol.* **240:** 372–384.

Khandogin, J. and Brooks III, C.L. 2006. Toward the accurate first-principles prediction of ionization equilibria in proteins. *Biochemistry* **45:** 9363–9373.

Kuramitsu, S. and Hamaguchi, K. 1980. Analysis of the acid-base titration curve of hen lysozyme. *Biochemistry* **87:** 1215–1219.

Labute, P. 2007 Protonate 3D: Assignment of macromolecular protonation state and geometry. Chemical Computing Group. http://www.chemcomp.com/journal/proton.htm

Laskowski Jr., M. and Sheraga, H.A. 1954. Thermodynamics considerations of protein reactions. I. Modified reactivity of polar groups. *J. Am. Chem. Soc.* **76:** 6305–6319.

Li, H., Robertson, A.D., and Jensen, J.H. 2005. Very fast empirical prediction and rationalization of protein pKa values. *Proteins* **61:** 704–721.

Li, X., Jacobson, M.P., Zhu, K., Zhao, S., and Friesner, R.A. 2007. Assignment of polar states for protein amino acid residues using an interaction cluster decomposition algorithm and its application to high resolution protein structure modeling. *Proteins* **66:** 824–837.

Linderstrom-Lang, K. 1924. On the ionization of proteins. *C.R. Trav. Lab. Carlsberg* **15:** 1–29.

Madura, J.D., Briggs, J.M., Wade, R.C., Davis, M.E., Luty, B.A., Ilin, A., Antosiewicz, J., Gilson, M.K., Bagheri, B., Scott, L.R., et al. 1995. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian dynamics program. *Comput. Phys. Commun.* **91:** 57–95.

Mehler, E.L. and Guarnieri, F. 1999. A self-consistent, microenvironment modulate descreened coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophys. J.* **77:** 3–22.

Momany, F. and Rone, R. 1992. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **13:** 888–900.

Mongan, J. and Case, D.A. 2005. Biomolecular simulations at constant pH. *Curr. Opin. Struct. Biol.* **15:** 157–163.

Mongan, J., Case, D.A., and McCammon, J.A. 2004. Constant pH molecular dynamics in generalized Born implicit solvent. *J. Comput. Chem.* **25:** 2038–2048.

Nielsen, J.E. and McCammon, J.A. 2003. On the evaluation and optimization of protein X-ray structures for pKa calculations. *Protein Sci.* **12:** 313–326.

Nielsen, J.E. and Vriend, G. 2001. Optimizing the hydrogen-bond network in Poisson–Boltzmann equation-based pKa calculations. *Proteins* **43:** 403–412.

Onufriev, A., Bashford, D., and Case, D.A. 2000. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **104:** 3712–3720.

Pace, C.N., Douglas, V., Laurents, D.V., and Thomson, J.A. 1990. pH dependence of the urea and guanidine hydrochloride denaturation of ribonuclease A and ribonuclease T1. *Biochemistry* **29:** 2564–2572.

Pfeil, W. and Privalov, P.L. 1976. Thermodynamic investigations of proteins. III. Thermodynamic description of lysozyme. *Biophys. Chem.* **4:** 41–50.

Pitera, J., Falta, M., and van Gunsteren, W. 2001. Dielectric properties of proteins from simulation: The effects of solvent, ligands, pH, and temperature. *Biophys. J.* **80:** 2546–2555.

Sampogna, R. and Honig, B. 1996. Electrostatic coupling between retinal isomerization and the ionization state of Glu204: A general mechanism for proton release in bacteriorhodopsin. *Biophys. J.* **71:** 1165–1171.

Schaefer, M., Sommer, M., and Karplus, M. 1997. pH-Dependence of protein stability: Absolute electrostatic free energy differences between conformations. *J. Phys. Chem. B* **101:** 1663–1683.

Schellman, J.A. 1975. Macromolecular binding. *Biopolymers* **14:** 999–1018.

Simonson, T. 2001. Macromolecular electrostatics: Continuum models and their growing pains. *Curr. Opin. Struct. Biol.* **11:** 243–252.

Smith, R., Breton, I.M., Chai, R.Y., and Kent, S.B.H. 1997. Ionization states of the catalytic residues in HIV-1 protease. *Nat. Struct. Biol.* **3:** 946–950.

Song, Y., Mao, J., and Gunner, M.R. 2003. Calculation of proton transfers in bacteriorhodopsin bR and M intermediates. *Biochemistry* **42:** 9875–9898.

Spassov, V.Z. and Bashford, D. 1999. Multiple-site ligand binding to flexible macromolecules: Separation of global and local conformational change and an iterative mobile clustering approach. *J. Comput. Chem.* **20:** 1091–1111.

Spassov, V.Z., Karshikoff, A.D., and Atanasov, A.P. 1989. Electrostatic interactions in proteins. A theoretical analysis of lysozyme ionization. *Biochim. Biophys. Acta* **999:** 1–6.

Spassov, V.Z., Luecke, H., Gerwert, K., and Bashford, D. 2001. p*K* calculations suggest storage of an excess proton in a hydrogen-bonded water network in bacteriorhodopsin. *J. Mol. Biol.* **312:** 203–219.

Spassov, V.Z., Yan, L., and Szalma, S.Z. 2002. Introducing an implicit membrane in Generalized Born/solvent accessibility continuum solvent models. *J. Phys. Chem. B* **106:** 8726–8738.

Spassov, V.Z., Yan, L., and Flook, P.K. 2007. The dominant role of side-chain backbone interactions in structural realization of amino acid code. ChiRotor: A side-chain prediction algorithm based on side-chain backbone interactions. *Protein Sci.* **16:** 494–506.

Still, W.C., Tempczyk, A., Hawley, R.C., and Hendrickson, T. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112:** 6127–6129.

Tan, Y.-J., Oliveberg, M., Davis, B., and Fersht, A.R. 1995. Perturbed pK_A-values in the denatured states of proteins. *J. Mol. Biol.* **254:** 980–992.

Tanford, C. and Kirkwood, J.G. 1957. Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Am. Chem. Soc.* **76:** 3331.

Tanford, C. and Roxby, R. 1972. Interpretation of protein titration curves. Application to lysozyme. *Biochemistry* **11:** 2192–2198.

Tanizaki, S. and Feig, M. 2006. Molecular dynamics simulations of large integral membrane proteins with an implicit membrane model. *J. Phys. Chem. B* **110:** 548–556.

Thurlkill, R.L., Grimsley, G.R., Scholtz, J.M., and Pace, C.N. 2006. pK values of the ionizable groups of proteins. *Protein Sci.* **15:** 1214–1218.

Todd, M.J., Semo, N., and Freire, E. 1998. The structural stability of the HIV-1 protease. *J. Mol. Biol.* **283:** 475–488.

Toseland, C.P., McSparron, H., Davies, M.N., and Flower, D.R. 2006. PPD v1.0—an integrated, web-accessible database of experimentally determined protein pK_a values. *Nucleic Acids Res.* **34:** D199–D203.

Trylska, J., Antosiewicz, J., Geller, M., Hodge, C.N., Klabe, R.M., Head, M.S., and Gilson, M.K. 1999. Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Sci.* **8:** 180–195.

Velazquez-Campoy, A., Luque, I., Todd, M.J., Milutinovich, M., Kiso, Y., and Freire, E. 2000. Thermodynamic dissection of the binding energetics of KNI-272, a potent HIV-1 protease inhibitor. *Protein Sci.* **9:** 1801–1809.

Wang, Y.X., Freedberg, D.I., Yamazaki, T., Wingfield, P.T., Stahl, S.J., Kaufman, J.D., Kiso, Y., and Torchia, D.A. 1996. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the asymmetric drug KNI-272. *Biochemistry* **35:** 9945–9950.

Warwicker, J. 2004. Improved pK_a calculations through flexibility based sampling of a water-dominated interaction scheme. *Protein Sci.* **13:** 2793–2805.

Word, J., Lovell, S., Richardson, J., and Richardson, D. 1999. Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285:** 1735–1747.

Yamazaki, T., Nicholson, L.K., Torchia, D.A., Wingfield, P., Stahl, S.J., Kaufman, J.D., Charles, J., Eyermann, C.J., Nicholas Hedge, C., Lam, P.Y.S., et al. 1994. Catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *J. Am. Chem. Soc.* **116:** 10791–10792.

Yang, A.S. and Honig, B. 1994. Structural origins of pH and ionic strength effects on protein stability. Acid denaturation of sperm whale apomyoglobin. *J. Mol. Biol.* **237:** 602–614.

Yang, A.S., Gunner, M.R., Sampogna, R., Sharp, K., and Honig, B. 1993. On the calculations of pK's in proteins. *Proteins* **15:** 252–265.

You, T.J. and Bashford, D. 1995. Conformation and hydrogen ion titration of proteins: A continuum electrostatic model with conformational flexibility. *Biophys. J.* **69:** 1721–1733.