PROTEINS: Structure, Function, and Genetics 43:403–412 (2001)

Optimizing the Hydrogen-Bond Network in Poisson–Boltzmann Equation-Based pK_a Calculations

Jens E. Nielsen* **and Gerrit Vriend**
European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

ABSTRACT **pK_a calculation methods that are based on finite difference solutions to the Poisson–Boltzmann equation (FDPB) require that energy calculations be performed for a large number of different protonation states of the protein. Normally, the differences between these protonation states are modeled by changing the charges on a few atoms, sometimes the differences are modeled by adding or removing hydrogens, and in a few cases the positions of these hydrogens are optimized locally. We present an FDPB-based pK_a calculation method in which the hydrogen-bond network is globally optimized for every single protonation state used. This global optimization gives a significant improvement in the accuracy of calculated pK_a values, especially for buried residues. It is also shown that large errors in calculated pK_a values are often due to structural artifacts induced by crystal packing. Optimization of the force fields and parameters used in pK_a calculations should therefore be performed with X-ray structures that are corrected for crystal artifacts. Proteins 2001;43:403–412.**

Key words: pK_a calculations; hydrogen-bond network optimization; electrostatics; Poisson–Boltzmann equation; protein structure optimization; crystal artifacts

INTRODUCTION

pK_a calculation methods have several applications in the field of protein chemistry. The most common ones are studies of the pH dependence of protein stability,[1] and of catalytic mechanisms of enzymes.[2,3] There has also been significant recent interest in incorporating pK_a calculations in molecular dynamics (MD) simulations.[4–6] The use of pK_a calculations for the above purposes has, however, been limited because of the poor agreement between calculations and experiments.[7] Consequently, there is strong interest in improving the accuracy of protein pK_a calculations.

The interest in calculating pK_a values dates back to the pioneering work of Linderstrøm-Lang,[8] who developed the theory for multiple titrating sites in a protein. Tanford and Roxby[9] augmented his work and successfully predicted the titration curve for lysozyme using a spherical model for the protein with all titratable groups on the surface. As more and more X-ray structures became available, it became evident that the titrational behavior of a protein had to be influenced both by its irregular shape and by the fact that not all titratable groups were situated at the surface. pK_a calculation techniques were at first not able to exploit this knowledge, since accurate electrostatic energies for a two-phase system could be calculated only for objects with a regular shape.[10,11] The development of a finite-difference Poisson–Boltzmann equation (FDPB) solver by Warwicker and Watson[12] made it possible to calculate electrostatic energies for an object of irregular shape. Bashford and Karplus[13] became the first to exploit this advantage when they constructed a pK_a calculation method based on an FDPB solver, and in subsequent years Yang et al.[14] and Antosiewicz et al.[7] reported similar methods. Some of these early FDPB-based methods experienced difficulties in beating the trivial null model,[7] which assumes that the pK_a values of the titratable groups in the protein are not shifted at all as compared with their model pK_a values (Table I). Many studies have consequently been undertaken to optimize the parameters of FDPB-based pK_a calculations in order to beat the null model.[15–19] The more successful methods are those of Demchuk and Wade,[19] and Antosiewicz et al.,[7] who adjusted the dielectric constants used in the calculation procedure.

Except for the work of Yang et al.,[14] no explicit attempt at modeling protein flexibility was made until You and Bashford[20] developed a theoretical basis for doing so. Recently, several implementations of protein flexibility in pK_a calculations have been published.[4–6,22–25] Many of these methods model the protein flexibility by applying molecular dynamics simulations. These methods do not, however, perform as well as the methods developed by Demchuk and Wade[19] and Antosiewicz et al.[7] Attempts to incorporate pK_a calculations in molecular dynamics (MD) simulations have been made,[4,5] but the improvements in the calculated pK_a values are modest.

The pK_a calculation methods that model protein flexibility explicitly are thus not yet able to compete with methods that use more empirical ways of describing the protein flexibility. It is obvious, however, that the fine details of protein flexibility play a key role in determining the pK_a values in the protein, and an accurate and explicit description of the protein flexibility must therefore improve the

Grant sponsor: Novo Nordisk A/S.

*Correspondence to: J.E. Nielsen, European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany. E-mail: nielsen@embl-heidelberg.de

Received 8 September 2000; Accepted 26 January 2001

TABLE I. Model pK_a Values Used in This Study

Residue type	Model pK_a value
Arginine	13.0
Aspartic acid	4.0
Cysteine	8.7
C-Terminus	3.8
Glutamic acid	4.4
Histidine	6.3
Lysine	10.4
N-Terminus	8.0
Tyrosine	9.6

accuracy of pK_a calculations. Ideally, it might be hoped that one could obtain precise pK_a values from a series of long MD simulations of a protein-bulk water system at different pH values. This is, however, not a feasible method with present-day computers. To improve the accuracy of protein pK_a calculation methods, we must construct a method that describes the important aspects of protein flexibility explicitly without performing long MD simulations. It has been shown that electrostatic field calculations are sensitive to the details of protein structures (see Sham et al.[24] for an excellent discussion of the importance of protein relaxation in protein pK_a calculations), and especially to changes in the hydrogen-bond network.[26] We shall show that the fine-tuning of the hydrogen-bond network used in FDPB-based pK_a calculations can lead to significant improvements in the accuracy of the calculated pK_a values.

FDPB-based pK_a calculations require the construction of a very large number of protein structures differing only in the protonation state of the titratable residues. Our method creates a globally hydrogen-bond optimized structure for each of these structures, and thus models the hydrogen-bond network of a particular protonation state as accurately as possible without using MD-based methods. The method models each protein protonation state by (1) globally optimizing the positions of the polar hydrogen atoms (this includes adding and removing titratable protons); and (2) allowing His, Asn, and Gln residues to "flip" (a 180-°rotation around their ξ^2, ξ^2, and ξ^3 angles, respectively). This model describes the flexibility of the hydrogen-bond network around each residue very accurately because the details of hydrogen-bond networks in proteins are well understood.[27–29]

X-ray Structures

An X-ray structure is normally perceived as a more "accurate" representation of a protein structure than a nuclear magnetic resonance (NMR) structure; consequently most pK_a calculation methods have been parameterized for use with X-ray structures. In the present study, we also use X-ray structures, but we are aware of the special problems that this poses: the presence of ions that co-crystallize with the protein and the effect of the crystal packing on the fine details of the protein surface.

pK_a Calculation Procedures

Several, sometimes very different, methods exist for calculating pK_a values. Methods that use the linear response approximation[24,25] Monte Carlo MD simulations,[30] self-consistent field theory,[31] simplified dielectric borders,[32] and sigmoidally screened electrostatic energies[33] have recently been published. The majority of these methods are not as accurate as the best FDPB-based methods, although the method of Mehler and Guarnieri[33] is a noteworthy exception.

Considerable efforts have been made to optimize the physical aspects of pK_a calculations,[4–7,13—25,34–36] such as dielectric shielding and protein flexibility. We describe here two very different ways of improving pK_a calculations that both relate to representing the protein structures: (1) a significant improvement in the accuracy of pK_a calculations can be obtained by globally optimizing the hydrogen-bond network for each structure needed in FDPB-based pK_a calculations, and (2) several of the largest errors in pK_a calculations are due to crystal induced structural artifacts.

We believe that the results presented in this article provide a good basis for redoing many of the earlier studies into the dielectric shielding and protein dynamics aspects of pK_a calculations. Combination of a rigorous optimization of these physical parameters with the structure optimization aspects discussed in this article will lead to even better pK_a calculation methods than we describe in this report.

MATERIALS AND METHODS

FDPB-based pK_a calculations methods compute the effect of transferring a titratable group from water to the appropriate position in the protein. The starting point for the calculation of every pK_a value is an estimate of the pK_a value in water for the residue in question. This estimate is called the model pK_a value and is determined by interpolating the pK_a values of compounds that resemble amino acid side chains (Table I). The pK_a value of a group in the protein is now computed by adding the desolvation energy and the interaction energy with nontitratable and titratable groups. The pK_a value of a titratable group in the protein, however, is not only dependent on its own environment, but also on the pK_a values of all other titratable groups in the protein and thereby dependent on the environment of those groups. This means that all pK_a values in the protein have to be computed simultaneously. This can be done by evaluating the Boltzmann sum shown below (eq. 1) for every titratable group at the pH values of interest:

$$\theta_i = \frac{\sum_S \delta_{i,S} e^{-E_S/kT}}{\sum_S e^{-E_S/kT}} \quad (1)$$

where θ_i is the fractional charge on group i, E_S is the energy of state S, and both sums are over all protonation states in the protein, with $\delta_{i,S}$ 1 only when group i is

charged in protonation state S and zero in all other cases. A protein with N titratable groups has 2^N protonation states, for which energies need to be calculated, which, of course, is not feasible for any but the smallest proteins. Consequently shortcuts have to be introduced to reduce the computation times. Most pK_a calculations use two methods to reduce the computation time. First, a Monte Carlo (MC) procedure is used to approximate the outcome of eq. 1 without summing over all the 2^N protonation states (see below). Second, the number of energy calculations needed to calculate the energies of all protonation states can be reduced by tabulating the interaction energies of pairs of titratable groups using eq 2:

$$G_S = \sum_i G_{\mathrm{pH},i} + \frac{1}{2} \sum_i \delta_i \sum_j \delta_j G_{i,j} \qquad (2)$$

where G_S is the energy of state S, $G_{\mathrm{pH},i}$ is a pH-dependent term for residue i that describes the effect of desolvation and the interaction with nontitratable charges, δ_i and δ_j are 1 only when their respective residues i and j are charged, and zero in all other cases. $G_{i,j}$ is the interaction energy between residues i and j. Although eq. 2 requires the calculation of only $\sim N^2$ pairwise interaction energies between titratable groups (plus $2N$ background interaction energies and desolvation energies), this still requires that the pK_a calculation algorithm models all the corresponding $\sim N^2$ structures. The methodological innovation that we describe in this article concerns the construction of these $\sim N^2$ structures.

The general pK_a calculation methodology has been described in detail elsewhere[14] and is not reviewed in this article. Details of the calculation method used in the present discussion can be found at http://www.cmbi.kun.nl/gv/nielsen/pK_a/calc_methodology.html.

Optimizing the Hydrogen-Bond Network in pK_a Calculations

In some pK_a calculation algorithms, hydrogen atoms are not modeled explicitly and only the charges and radii of heavy atoms are modified. In other methods hydrogen atoms are removed or added in standard positions,[14,15] and in a two cases multiple locally optimized positions for each hydrogen atom were used,[22,34] resulting in a significant improvement in the accuracy of the calculated pK_a values for hen egg white lysozyme (HEWL). We optimize the hydrogen-bond network globally for each of structures needed in FDPB-based pK_a calculations using the method of Hooft et al.[27] This method additionally allows for His, Asn, and Gln side chain flips if these would produce a better hydrogen-bond network. Using a global rather than a local optimization protocol is important because hydrogen bonds are normally part of large networks, so that a hydrogen bond far away from a titratable group can influence the hydrogen positions on nearby residues. When constructing hydrogen atom positions (and especially when choosing the titratable atom of His, Asp, and Glu residues) one therefore has to optimize the energy of the entire hydrogen-bond network and not only the energy of each individual hydrogen bond.[27] The importance of this is illustrated in Figure 1, where the titratable atom of the His can be either the Nδ1 or the Nϵ2, depending on how the hydrogen-bond energy is calculated.

Methods for Constructing Hydrogen Positions

We have evaluated two different protocols for constructing hydrogen-atom positions in pK_a calculations. These are described below.

Standard coordinate method (standard)

The standard method constructs hydrogen atoms on every residue without considering the local environment. Hydrogen atoms are constructed according to the distances and angles specified in the WHAT IF[37] topology file, and there is no optimization of hydrogen bonds. Histidines always titrate at their Nϵ2, and Glu and Asp protons are bound to the Oϵ2 and Oδ2 atoms, respectively.

Global hydrogen-bond network optimization (H-bond Opt.)

The hydrogen-bond optimization procedure by Hooft et al.[27] as implemented in WHAT IF is used to construct the hydrogen-bond network. The method of Hooft et al. method uses a statistically derived force field to calculate the energy of every single hydrogen bond. The different protonation states needed in FDPB-based pK_a calculations are constructed by finding the globally most favorable hydrogen-bond network for each protonation state. This means that at least two structures are created for every titratable group (four structures are created for strongly interacting ion pairs).

Exceptions are the protonation states of the terminal groups and the protonation states of Arg and Lys residues. In the case of Lys and N-terminal residues, this is of minor importance as the difference in the hydrogen-bonding capabilities of an NH_2 group and an NH_3 group are minimal.[27] The deprotonation of Lys and N-terminal residues is therefore performed by removing the proton with the lowest hydrogen-bonding energy. The deprotonated state of arginine residues was modeled by changing the partial charges of the atoms in the residue (Table II), and the protonated form of the C-terminal group was constructed by adding a proton to the oxygen that could form the best possible hydrogen bond. The chance that this will cause inconsistencies in the hydrogen-bond network is very small, as very few proteins are folded at the pH values where arginine residues titrate, and C-terminal residues are normally located near the protein surface.

Solving the Poisson–Boltzmann Equation

The program DelPhi II[38] was used to calculate the electrostatic energies. Desolvation energies were calculated in a single run using a 65 cubed grid with a resolution of 3.0 grid points/Å[14] because trials indicated that the gain in accuracy using focusing runs was insignificant. Background interaction energies and the pairwise interaction energies between titratable groups were calculated using four focusing runs on a 65-cubed grid. The initial grid had a fill percentage of 30, and the final grid resolution was 4.0

grid points/Å. Background interaction energies are always calculated from the 4.0-grid point/Å map, whereas the pairwise interaction energies for titratable groups were calculated from the highest resolution potential map containing both groups in the pair. The interaction energies for the charged–neutral, neutral–charged, and neutral–neutral states for a pair of titratable groups make a significant contribution to the overall interaction energy of the pair only if the charged–charged interaction is strong.[14]

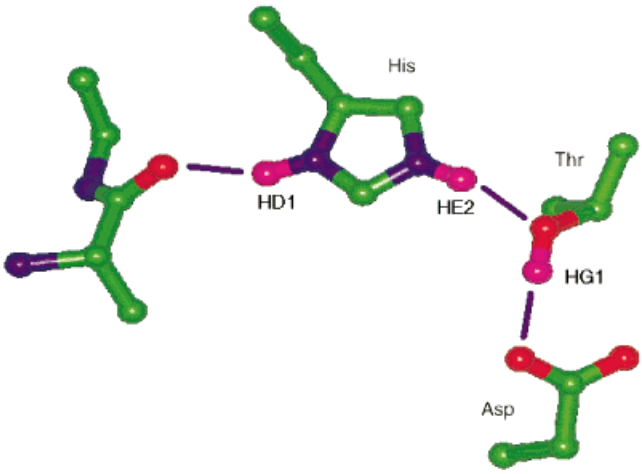

Fig. 1. Hypothetical Thr, His, Asp system. Hydrogen bonds are shown as blue lines. The His Nδ1 hydrogen bond is significantly weaker than the Nε2 hydrogen bond as measured by the standard distance and angle criteria for hydrogen bonds (Hooft et al., 1996), and the Nδ1 proton is therefore predicted to titrate first if the hydrogen-bond energy is measured only locally. If the global hydrogen-bond energy is used, it is seen that the Nε2-proton will titrate, as the proton on the threonine can compensate, at least partially, for the lost Nε2 hydrogen bond.

These energies were therefore calculated only if the charged–charged interaction energy was ≥1.0 kT. In all other cases, they were assumed to be zero.[14] The OPLS force field[39] was used as source of charges and radii. The parameters for the nonstandard protonation forms of Glu, Asp, Tyr, Cys, Arg, and Lys are shown in Table II. Two values were used for the protein dielectric constant. The electrostatic energy calculations were performed with a protein dielectric constant of 16 for residues that were involved in crystal contacts, residues with a B-factor of >30 and residues with a second accessible rotamer. The electrostatic energy calculations for all remaining residues were performed with a protein dielectric constant of 8. The dielectric constant of the solvent was in all cases set to 80. Pairwise interaction energies were calculated with a dielectric constant of 16 if one of the two groups in the pair fulfilled the criteria for calculation with a dielectric constant of 16. All other pairwise interaction energies were calculated with a dielectric constant of 8. Residues with an accessible alternative rotamer were identified using the WHAT IF[37] position specific rotamer libraries.[40]

Calculating pK_a Values

Theories for calculating pK_a values using Poisson–Boltzmann solvers have been reviewed elsewhere.[13,14] Our pK_a calculation method generally follows the scheme of Yang et al.,[14] with the exception that titration curves are determined using Monte Carlo sampling.[41] The system was allowed to equilibrate for 10,000 steps before the sampling proceeded for 180,000 steps. The degree of protonation for each titratable group was calculated from pH 0.0 to pH 20.0 in steps of 0.1 pH unit. pK_a values were in all cases defined as the pH values in which the titratable

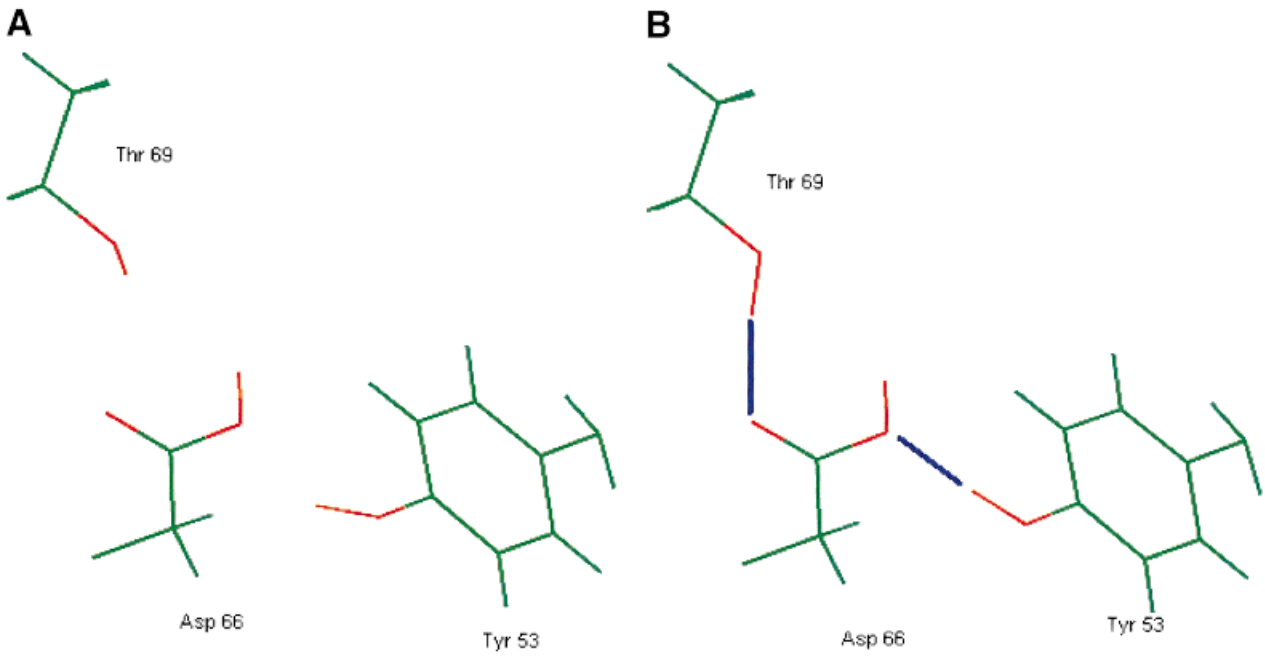

Fig. 2. The residues used in the calculation of the desolvation and background interaction energies for HEWL Asp 66. The only difference between neutral and charged structures for the std-coor. and the stat-GN calculations is the absence/presence of the proton on Asp 66, and consequently only the structure with the neutral Asp is shown. Hydrogen bonds are shown as blue lines. **A:** standard method, Asp 66 neutral; **B:** H-bond Opt., Asp 66 neutral.

TABLE II. Charge and Radius Parameters for the Nonstandard Protonation Forms of Asp, Glu, Tyr, Arg, and Lys*

	Neutral Asp		Neutral Glu		Charged Tyr		Neutral Arg		Neutral Lys	
Atom	Charge	rad	Charge	rad	Charge	rad	Charge	rad	Charge	rad
N	−0.570	1.625	−0.570	1.625	−0.570	1.625	−0.570	1.625	−0.570	1.625
HN	0.370	1.200	0.370	1.200	0.370	1.200	0.370	1.200	0.370	1.200
CA	0.200	1.900	0.200	1.900	0.200	1.900	0.200	1.900	0.200	1.900
C	0.500	1.875	0.500	1.875	0.500	1.875	0.500	1.875	0.500	1.875
O	−0.500	1.480	−0.500	1.480	−0.500	1.480	−0.500	1.480	−0.500	1.480
CB	0.000	1.952	0.000	1.952	0.000	1.952	0.000	1.952	0.000	1.952
CG	0.800	1.875	0.000	1.952	0.000	1.875	0.000	1.952	0.000	1.952
OD1	−0.685	1.480	—	—	—	—	—	—	—	—
OD2	−0.485	1.480	—	—	—	—	—	—	—	—
HD2	0.370	1.200	—	—	—	—	—	—	—	—
CD[a]	—	—	0.800	1.875	0.000	1.875	0.200	1.952	0.000	1.952
OE1	—	—	−0.685	1.480	—	—	—	—	—	—
OE2	—	—	−0.485	1.480	—	—	—	—	—	—
HE2	—	—	0.370	1.200	—	—	—	—	—	—
CE[b]	—	—	—	—	0.000	1.875	—	—	0.000	1.952
CZ	—	—	—	—	−0.221	1.875	0.020	1.125	—	—
OH	—	—	—	—	−0.429	1.535	—	—	—	—
HE	—	—	—	—	—	—	0.350	1.200	—	—
NE	—	—	—	—	—	—	−0.570	1.625	—	—
NH1/NH2	—	—	—	—	—	—	−0.440	1.625	—	—
Arg HH[c]	—	—	—	—	—	—	0.220	1.200	—	—
NZ	—	—	—	—	—	—	—	—	−0.300	1.625
HZ1/HZ2	—	—	—	—	—	—	—	—	0.150	1.200

rad, radius (in Å).
*All nonpolar protons were assigned zero charge and zero radius. The neutral state of Cys was modeled by the parameters for a cysteine in a disulfide bridge.
[a]For Tyr, both CD1 and CD2.
[b]For Tyr, both CE1 and CE2.
[c]Protons 1HH1, 1HH2, 2HH1, and 2HH2.

group was half-protonated. pK_a values were calculated for Asp, Glu, Tyr, His, Cys, Arg, and Lys residues.

X-Ray Structures

Calculations were carried out for the 3D structures of hen egg white lysozyme (HEWL), bovine pancreatic trypsin inhibitor (BPTI), ribonuclease A (RNase A), *Bacillus circulans* xylanase (Xylanase), calbindin, protein G (PG), third domain of turkey ovomucoid inhibitor (OMTKY3), ribonuclease H (RNase H), and barnase. Structures of all these molecules are available from the PDB.[42] Structures with the following accession codes were used: HEWL: 2lzt,[43] BPTI: 4pti,[44] RNase A: 3rn3,[45] Xylanase: 1xnb,[46] Calbindin: 3icb,[47] PG: 1pga,[48] OMTKY3: 1ppf,[49] RNase H: 2rn2,[50] Barnase: 1a2p.[51]

Ions that are considered to be crystallization artifacts are removed from the X-ray structures to try to model the solution structure of the protein more closely. The calculations are performed without crystal waters.

RESULTS

We examined the effects on protein pK_a calculations of two methods for constructing hydrogen positions: (1) the standard coordinate method (standard), and (2) the global network method (H-bond Opt.). In both cases, we report the accuracy of the calculated pK_a values as the root-mean-square difference (RMSD) between the calculated and experimentally measured pK_a values. (A list of the experimental and calculated pK_a values and the comparisons used in this study are available at http://www.cmbi.kun.nl/gv/nielsen/pKa/results.)

Optimizing the Hydrogen-Bond Network

Generally, the pK_a values obtained with both methods are in good agreement with experimental results. The overall RMSD for all titratable groups is 0.93 and 0.91 for the standard and H-bond Opt. Methods, respectively. The percentage of pK_a values within 0.5 unit of the experimental value is standard: 55 and H-bond Opt. 49. Both methods give pK_a values for more than 10 residues, which are in error by more than 1.5 units. The reasons for these large errors are discussed below.

Table III compares the accuracy of the calculated pK_a values for nine proteins. It can be seen that the H-bond Opt. method improves the accuracy for the majority of the proteins compared with the standard method, but for two proteins, the results are similar. A detailed comparison of the differences in pK_a values between the standard and the H-bond Opt. methods shows that for some groups, the calculated pK_a values become less accurate. This is the case for His 12 in RNase A. Inspection of the hydrogen-bond network around this residue shows that the hydrogen-bond network is better in the H-bond Opt. method than in the standard method. There are good reasons, however, for

TABLE III. Root-Mean-Square Deviations (RMSD) Between Experimental Values and Values Calculated With the Standard and Hydrogen-Bond Optimized Methods*

Molecule PDB code	HEWL 2lzt	BPTI 4pti	RNase A 3rn3	Barnase 1a2p	Protein G 1pga	RNase H 2rn2	Xylanase 1xnb	OVO 1ppf	Calbindin 3icb
H-bond Opt.	0.66	0.66	1.37	0.90	0.87	1.08	1.18	0.59	0.42
Standard	0.74	0.66	1.52	1.07	0.72	0.87	1.32	0.65	0.40
RMSD (H-bond Opt.-standard)	−0.12	0.00	−0.12	−0.17	0.15	0.21	−0.24	−0.06	0.02
No. of groups with known pK_a value	21	14	16	13	13	25	12	6	19

*Top two rows: RMSD values for both methods. Row 3: RMSD improvements when going from the standard method to the hydrogen-bond optimized method.

TABLE IV. RMSD Values for Hen Egg White Lysozyme (HEWL) and Bovine Pancreatic Trypsin Inhibitor (BPTI) Obtained With Poisson–Bolztmann Equation-Based pK_a Calculation Methods

Method	HEWL[a]	BPTI	Xylanase[a]
Null model	1.04 (−)	0.67	—
Yang et al. (1993)[14]	1.07 (−)	—	—
Antosiewicz et al. (1994)[7]	0.65 (−)	0.44	—
You and Bashford (1995)[20]	1.87 (−)	—	—
Antosiewicz et al. (1996)[16]	2.12 (+)	—	—
Demchuk and Wade (1996)[19]	0.65 (−)	0.62	—
Alexov and Gunner (1997)[22]	1.52 (−)	—	—
WHAT IF (standard)	0.74 (−)	0.66	1.32 (−)
WHAT IF (H-bond Opt.)	0.66 (+)	0.66	1.18 (+)

[a]+ sign in parentheses (+) indicates whether the active site residues (HEWL: Glu 35 and Asp 52, xylanase: Glu 172 and Glu 78) were predicted so that the mechanistic roles of the two groups could be inferred from the pK_a values. For HEWL, this was defined as pK_a (Glu 35)-pK_a (Asp 52) ≥ 1.5 units, and pK_a (Glu 35) ≥ 5.0, for xylanase as pK_a (Glu 172)-pK_a (Glu 78) ≥ 1.5).

the decrease in accuracy. In the crystal structure of RNase A, a sulfate ion is found within a hydrogen-bonding distance of His 12. This sulfate ion probably perturbs the structure around His 12; optimization of the hydrogen-bond network in this region is therefore influenced by the perturbations the sulfate ion makes on the structure (although the sulfate ion itself is not included in the calculations). Some interesting cases showing successes and failures of the H-bond Opt. method are discussed below.

Hen Egg White Lysozyme

Table IV compares the performance of some commonly used pK_a calculation algorithms for HEWL. It can be seen that the H-bond Opt. method is the only FDPB-based method capable of producing a low overall RMSD, while predicting the pK_a values of the active site residues correctly. Furthermore, the H-bond Opt. method is capable of predicting the pK_a value for Asp 66, which has a low pK_a value despite the fact that it is completely buried. Asp 66 is generally predicted wrongly by most pK_a calculation packages, and it is therefore remarkable that the H-bond Opt. method performs so well for this residue. An even better example of the advantages of the H-bond Opt. algorithm is provided by His 15, where there full potential of the H-bond Opt. method is revealed.

Asp 66

The experimental pK_a value for Asp 66 is 2.0. The standard method calculates it as 3.6, while the H-bond Opt. method calculates it as 2.2. The hydrogen-bond network around Asp 66 used in the standard and H-bond Opt. calculations is shown in Figure 2. The network is shown only for the protonated state of Asp 66, since the only difference in the hydrogen-bond networks between the charged and neutral form for Asp 66 is the titratable proton on the Asp. In the standard method, the protons on Thr 69 and Tyr 53 do not form a hydrogen bond to Asp 66. The interaction energy between the surroundings and the charged state of Asp 66 is therefore less favorable; consequently, a higher pK_a value is calculated for Asp 66. Using the H-bond Opt. method, the hydrogen bonds from Thr 69 and Tyr 53 to Asp 66 are in place. Both of these hydrogen bonds create a more favorable interaction between the surroundings and the charged state of Asp 66, so that a lower pK_a value is calculated.

His 15

The pK_a of His 15 is 5.7. The standard and H-bond Opt. methods give 5.3 and 4.9, respectively, which are both very close to the experimental value. The hydrogen-bond networks used for the calculations are shown in Figure 3. It can be seen that it is the Nϵ2 hydrogen of His 15 that titrates because the proton on Thr 89 is able to substitute for the lost hydrogen bond. At the same time, the proton on Asp 87 shifts from Oδ1 to Oδ2 and is thus able to hydrogen bond to Thr 89. The results for His 15 do not improve when the hydrogen bonds are optimized, even though the hydrogen-bond network around His 15 is undoubtedly better in the H-bond Opt. calculations than in the standard calculations.

Active Site pK_a Values

A major use of pK_a calculation programs is to estimate pK_a values of active site residues. These are, of course,

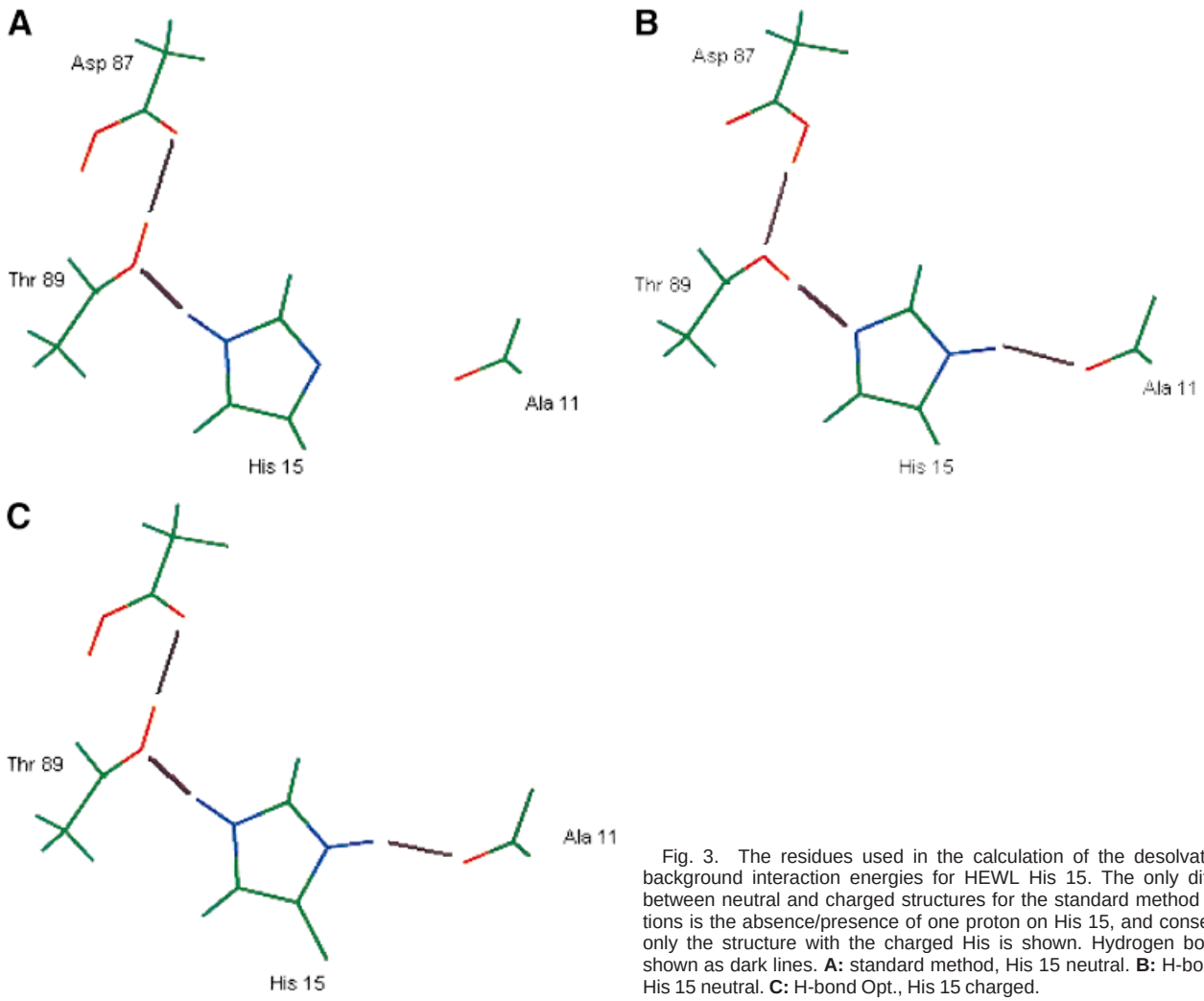

Fig. 3. The residues used in the calculation of the desolvation and background interaction energies for HEWL His 15. The only difference between neutral and charged structures for the standard method calculations is the absence/presence of one proton on His 15, and consequently only the structure with the charged His is shown. Hydrogen bonds are shown as dark lines. **A:** standard method, His 15 neutral. **B:** H-bond Opt., His 15 neutral. **C:** H-bond Opt., His 15 charged.

TABLE V. RMSD Between Calculated and Experimental pK_a Values for Exposed and Buried Groups in the Test Set*

	8 Buried groups	116 Exposed groups	All groups
H-bond Opt.	1.25	0.90	0.91
Standard	2.17	0.79	0.93

*Buried groups were defined as residues where the average solvent-exposed surface of the titratable atoms was <0.75 Å^2. Exposed groups are defined as the remaining groups.

more problematic than all other pK_a values, and several pK_a calculation programs are poor at calculating the correct pK_a values for these residues. The H-bond Opt. method, however, performs quite well on the active site pK_a values of HEWL and *Bacillus circulans* xylanase (which are the only two enzymes whose active site pK_a values are known). In both cases the H-bond Opt. algorithm makes it possible to assign functional roles to the two key active site residues (Table IV), although the absolute values are not quite correct for *Bacillus circulans* xylanase.

Sources of Error

The results for His 15 in HEWL and for His 12 in RNase A makes it clear that optimizing the hydrogen-bond network does not always improve the agreement between calculated and experimentally observed pK_a values. Visual inspection suggests that the agreement is worse for the residues in regions involved in symmetry contacts or those that are likely to be flexible.

Flexible regions

The influence of protein flexibility is illustrated by His 119 in RNase A, which sits in the active site and is known to have a pK_a value of 6.10. This residue is found in two rotamers in the Protein Data Bank (PDB) file; it is therefore likely that it is highly mobile in the free enzyme. The calculated pK_a value for His 119 is too low for both rotamers. This happens when either the standard or the H-bond Opt. method is used. In fact, the pK_a value becomes even more inaccurate for one rotamer when using is H-bond Opt. method instead of the standard method. This is not unexpected, as the measured pK_a value of His 119 is a weighted average of the pK_a values of all the

TABLE VI. Residues Incorrect by >1.5 Units

Residue	Standard	H-bond Opt.	Location
BPTI–Tyr 35	1.8	1.8	Partly buried; no symmetry contacts
Protein G–Glu 27	1.9	2.4	Surface salt bridge with Lys 28
RNase A–Glu 2	2.3	2.4	Surface salt bridge with Arg 10
RNase A–His 12	0.7	3.9	Sulfate ion
RNase A–Asp 14	2.4	0.4	Close to Arg 33—on the surface
RNase A–His 48	4.1	2.3	Close to Asp 14
RNase A–Asp 121	2.5	1.1	Close to Lys 66, which is influenced by crystal packing
RNase H–Glu 48	1.2	2.2	3.8 Å from a symmetry-related His 83
RNase H–Glu 57	1.7	2.1	Surface salt bridge with Arg 106
RNase H–Glu 119	1.6	2.0	Surface salt bridge with His 127
RNase H–Glu 129	1.7	2.0	Surface salt bridge with Arg 27
Lysozyme–Glu 35	1.4	0.6	Active site
Lysozyme–Asp 52	1.1	1.9	Active site
Lysozyme–Asp 66	1.6	0.2	Buried
Xylanase–Asp 83	2.3	>2.0	H-bonds to Arg 135 (on the surface)
Xylanase–Asp 101	0.9	1.6	Close to His 149, buried water molecule
Xylanase–His 149	>2.0	1.6	Buried water molecule
Xylanase–Glu 172	2.4	1.7	Active site
Barnase–Asp 52	>2.2	1.86	Surface salt bridge with Lys 27

conformations that it occupies. This includes conformations between the two rotamers in the PDB file, and it is therefore obvious that optimizing the hydrogen-bond network for either of the two rotamers found in the PDB file cannot be expected to improve the agreement between the calculated and experimentally observed pK_a value.

Table V shows the RMSD for the exposed and buried groups in the test set of nine proteins. The RMSD for the deeply buried groups decreases significantly when the H-bond Opt. methods is used, whereas the RMSD for the other groups increases slightly. This shows that hydrogen-bond optimization works when the structure is rigid and well defined, as in the protein interior, but for flexible residues at the surface it is less likely that optimizing the hydrogen-bond network will improve the pK_a values. There is also an influence from crystal packing, which affects surface residues predominantly.

Effect of crystal packing

Both methods made errors of more than 1.5 units for more than 10 residues. Table VI shows a summary of the errors in the calculated pK_a values for these residues and provides an overview of their location. Of the 19 residues with pK_a values out by more than 1.5 units, eight are involved in surface salt bridges, four are influenced by changes in the protein structure induced by the crystal packing, and two are close to a completely buried water molecule that was not included in the calculations. One additional group is in error only in the standard method (Asp 66 in HEWL), but for the remaining four groups we could find no obvious explanation for the poor agreement with experimental values. It is, however, clear from the summary above that both methods are very bad at predicting pK_a values for residues involved in surface salt bridges. This is probably because surface salt bridges are present only transiently when the protein is in solution, so it is not surprising that there is a discrepancy between the measured pK_a values (from NMR) and the calculated pK_a values (using an X-ray structure). Ideally, we should therefore use an ensemble of structures for the surface residues to model their titrational behavior correctly.

The present method corrects for the influence of flexibility and crystal contacts by employing a dielectric constant of 16 in the electrostatic energy calculations for a subset of the titratable groups (as mentioned in the materials and methods section). Electrostatic energy calculations for a residue are carried out with a dielectric constant of 16 when the residue fulfills at least one of the following criteria: (1) the average B-factor of the residue is above 30, (2) the residue has an accessible alternative rotamer, and (3) the residue is involved in crystal contacts.

The use of a protein dielectric constant of 16 for residues that fulfill these criteria improves the overall RMSD between the calculated and experimental values from 1.11 to 0.91. Much more can, of course, be done to correct an X-ray structure for the effects of flexibility and crystal packing: MD simulations, remodeling of the protein surface, and the use of NMR structures. We have used the above strategy of employing two different dielectric constants because it is simple and easy to implement. In this way, we reduce the influence of the local environment on the pK_a values in areas for which the X-ray structure is likely to deviate from the solution structure of the protein. Work on alternative methods to make an X-ray structure more solution-like is in progress.

DISCUSSION

We have shown that optimizing the hydrogen-bond networks in FDPB-based pK_a calculations improves the agreement with experimental values for buried titratable groups (Table V). Using two different values for the protein dielectric constants also improved the RMSD.

The method calculates the pK_a values for Asp 66, Glu 35, and Asp 52 in HEWL correctly, which has never before

been achieved with FDPB-based pK_a calculations. Asp 52 and Glu 35 are the catalytic nucleophil and hydrogen donor, respectively. Getting these two pK_a values right justifies the use of pK_a calculations for identifying the roles of active site residues. The pK_a value of Asp 66 is especially difficult to calculate because Asp 66 is completely buried, and it still comes out with a very low pK_a value. The reason for the low pK_a value is a very favorable local hydrogen-bond network, and pK_a calculation methods that disregard the details of the hydrogen-bond network are bound to calculate a wrong pK_a for this residue.

The pK_a values calculated by the H-bond Opt. method allow for the identification of the hydrogen donor and the nucleophil in the *Bacillus circulans* xylanase, and the method thus calculates all four active site pK_a values in the test set correctly. Mehler and Guarnieri[33] recently published a remarkably successful pK_a calculation method, that provides a significant improvement in the accuracy of protein pK_a calculations. The method uses a sigmoidal screening of the electrostatic energies, modulated by the hydrophobicity of the environment of each titratable residue. It is questionable, however, whether such a screening function based only on the hydrophobicity of the local environment can describe all the effects of the very complex dynamics of the hydrogen-bond networks in proteins.

The results of pK_a calculation algorithms are always compared with pK_a values measured by solution state NMR experiments, but normally X-ray structures, and not NMR structures, are used as source of coordinates for pK_a calculation algorithms. X-ray protein structures are known to be slightly different from protein solution structures, especially in the regions in which protein atoms make symmetry-related contacts. In the development of pK_a calculation methods, this has been largely ignored. The parameters of most pK_a calculation algorithms have thus been optimized to give the best agreement with experimental data, regardless of the differences between the solution and the crystal state of the protein. This obviously leads to an inconsistent parameter set and, in the worst-case scenario, it will lead to the situation that residues involved in crystal contacts will have their pK_a values predicted more correctly than residues that have identical conformations both in the crystal and in solution.

Table VI shows that some of the largest errors with our method occur for residues that are involved in crystal contacts. Obviously, correcting for crystal artifacts in these cases will lead to improvements. It is, however, very likely that a much better overall performance of pK_a calculation algorithms can be obtained by correcting the parameters and the force fields for the errors that have been introduced by calibrating the algorithms on X-ray structures with crystallization artifacts.

ACKNOWLEDGMENTS

The authors thank Rob Hooft for help with the hydrogen-bond optimization algorithm, Rebecca Wade for helpful discussions and Barry Honig for making DelPhi II available.

REFERENCES

1. Tanford C. Protein denaturation. Part C. Theoretical models for the mechanism of denaturation. Adv Protein Chem 1970;25:1–95.
2. Lamotte-Brasseur J, Lounnas V, Raquet X, Wade RC. pK_a calculations for class A beta-lactamases: influence of substrate binding. Protein Sci 1999;8:404–409.
3. Raquet X, Lounnas V, Lamotte-Brasseur J, Frere JM, Wade RC. pK_a calculations for class A beta-lactamases: methodological and mechanistic implications. Biophys J 1997;73:2416–2426.
4. Sandberg L, Edholm O. A fast and simple method to calculate protonation states in proteins. Proteins 1999;36:474–483.
5. Baptista AM, Martel PJ, Petersen SB. Simulation of protein conformational freedom as a function of pH: constant-pH molecular dynamics using implicit titration. Proteins 1997;27:523–544.
6. Van Vlijmen, Schaefer M, Karplus M. Improving the accuracy of protein pK_a calculations: conformational averaging versus the average structure. Proteins 1998;33:145–158.
7. Antosiewicz J, McCammon JA, Gilson MK. Prediction of pH-dependent properties of proteins. J Mol Biol 1994;238:415–436.
8. Linderstrøm-Lang K. The ionisation of proteins. CR Trav Lab Carlsberg 1924;15:1–29.
9. Tanford C, Roxby R. Interpretation of protein titration curves. Application to lysozyme. Biochemistry 1972;11:2193–2198.
10. Matthew JB. Electrostatic effects in proteins. Annu Rev Biophys Biophys Chem 1985;14:387–417.
11. Klapper I, Hagstrom R, Fine R, Sharp K, Honig B. Focusing of electric fields in the active site of Cu–Zn superoxide dismutase: effects of ionic strength and amino-acid modification. Proteins 1986;1:47–59.
12. Warwicker J, Watson HC. Calculation of the electric potential in the active site cleft due to alpha-helix dipoles. J Mol Biol 1982;157: 671–679.
13. Bashford D, Karplus M. pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29:10219–10225.
14. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of pK_as in proteins. Proteins 1993;15:252–265.
15. Antosiewicz J, Briggs JM, Elcock AH, Gilson MK, McCammon JA. Computing ionisation states of proteins with a detailed charge model. J Comput Chem 1996;17:1633–1644.
16. Antosiewicz J, McCammon JA, Gilson MK. The determinants of pK_as in proteins. Biochemistry 1996;35:7819–7833.
17. Karshikoff A. A simple algorithm for the calculation of multiple site titration curves. Protein Eng 1995;8:243–248.
18. Gibas CJ, Subramaniam S. Explicit solvent models in protein pK_a calculations. Biophys J 1996;71:138–147.
19. Demchuk E, Wade RC. Improving the continuum dielectric approach to calculating pK_as of ionizable groups in proteins. J Phys Chem 1996;100:17373–17387.
20. You TJ, Bashford D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. Biophys J 1995;69:1721–1733.
21. Beroza P, Case DA. Including side chain flexibility in continuum electrostatic calculations of protein titration. J Phys Chem 1996; 100:20156–20163.
22. Alexov EG, Gunner MR. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. Biophys J 1997;72:2075–2093.
23. Zhou HX, Vijayakumar M. Modeling of protein conformational fluctuations in pK_a predictions. J Mol Biol 1997;267:1002–1011.
24. Sham YY, Chu ZT, Warshel A. Consistent calculations of pK_as of ionizable residues in proteins: semi-microscopic and microscopic approaches. J Phys Chem 1997;101:4458–4472.
25. Sham YY, Muegge I, Warshel A. The effect of protein relaxation on charge–charge interactions and dielectric constants of proteins. Biophys J 1998;74:1744–1753.
26. Nielsen JE, Andersen KV, Honig B, Hooft RW, Klebe G, Vriend G, Wade RC. Improving macromolecular electrostatics calculations. Protein Eng 1999;12:657–662.
27. Hooft RW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 1996;26:363–376.
28. McDonald IK, Thornton JM. The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. Protein Eng 1995;8:217–224.
29. Bass MB, Hopkins DF, Jaquysh WA, Ornstein RL. A method for

determining the positions of polar hydrogens added to a protein structure that maximizes protein hydrogen bonding. Proteins 1992;12:266–277.
30. Kesvatera T, Jonsson B, Thulin E, Linse S. Ionization behavior of acidic residues in calbindin D(9k). Proteins 1999;37:106–115.
31. Dimitrov RA, Crichton RR. Self-consistent field approach to protein structure and stability. I. pH dependence of electrostatic contribution. Proteins 1997;27:576–596.
32. Warwicker J. Simplified methods for pK_a and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. Protein Sci 1999;8:418–425.
33. Mehler EL, Guarnieri F. A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. Biophys J 1999;77:3–22.
34. Spassov VZ, Bashford D. Multiple-site ligand binding to flexible macromolecules: separation of global and local conformational change and an iterative mobile clustering approach. J Comp Chem 1999;20:1091–1111.
35. Alexov EG, Gunner MR. Calculated protein and proton motions coupled to electron transfer: electron transfer from QA- to QB in bacterial photosynthetic reaction centers. Biochemistry 1999;38: 8253–8270.
36. Havranek JJ, Harbury PB. Tanford–Kirkwood electrostatics for protein modeling. Proc Natl Acad Sci USA 1999;96:11145–11150.
37. Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph 1990;8:52–56.
38. Nicholls A, Honig B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. J Comp Chem 1991;12:435–445.
39. Jorgensen WL, Tirado-Rives J. The OPLS potential functions for proteins: energy minimizations for crystals for cyclic peptides and crambin. J Am Chem Soc 1988;110:1657–1666.
40. Chinea G, Padron G, Hooft RW, Sander C, Vriend G. The use of position-specific rotamers in model building by homology. Proteins 1995;23:415–421.
41. Beroza P, Fredkin DR, Okamura MY, Feher G. 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 USA 1991;88:5804–5808.
42. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research 2000;28:235–242.
43. Ramanadham M, Sieker LC, Jensen LH. Refinement of triclinic lysozyme: II. The method of stereochemically restrained least squares. Acta Crystallogr B 1990;46:63–69.
44. Marquart M, Walter J, Deisenhofer J, Bode W, Huber R. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors Acta Crystallogr B 1983;39:480–485.
45. Howlin B, Moss DS, Harris GW. Segmented anisotropic refinement of bovine ribonuclease A by the application of the rigid-body TLS model. Acta Crystallogr A 1989;45:851–861.
46. Wakarchuk WW, Campbell RL, Sung WL, Davoodi J, Yaguchi M. Mutational and crystallographic analyses of the active site residues of the *Bacillus circulans* xylanase. Protein Sci 1994;3:467–475.
47. Szebenyi DM, Moffat K. The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. J Biol Chem 1986;261:8761–8777.
48. Gallagher T, Alexander P, Bryan P, Gilliland GL. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. Biochemistry 1994;33: 4721–4729.
49. Bode W, Wei AZ, Huber R, Meyer E, Travis J, Neumann S. X-ray crystal structure of the complex of human leukocyte elastase (PMN elastase) and the third domain of the turkey ovomucoid inhibitor. EMBO J 1986;5:2453–2458.
50. Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Nakamura H, Ikehara M, Matsuzaki T, Morikawa K. Structural details of ribonuclease H from *Escherichia coli* as refined to an atomic resolution. J Mol Biol 1992;223:1029–1052.
51. Mauguen Y, Hartley RW, Dodson EJ, Dodson GG, Bricogne G, Chothia C, Jack A. Molecular structure of a new family of ribonucleases. Nature 1982;297:162–164.