Some Theoretical and Computational Aspects of the Inclusion of Proton Isomerism in the Protonation Equilibrium of Proteins

António M. Baptista* and Cláudio M. Soares

Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa, 2781-901 Oeiras, Portugal

Received: August 2, 2000

The present article discusses some aspects concerning the inclusion of proton isomerism in simulations of the global protonation equilibrium of protein molecules. In the context of continuum electrostatic methods, the usual basis for these simulations, this isomerism can be treated as a coexistence of tautomeric forms in equilibrium in a rigid structure; furthermore, it can be formally extended to nontitrable sites with proton isomerism, such as alcohol groups and water molecules. We follow the previously adopted approach of transforming the real system of tautomeric sites into a thermodynamically equivalent one of nontautomeric pseudosites, establishing a proper relation between the two systems. The necessary energetic and entropic modifications of model compound pK_a values are also discussed. Additionally, we discuss the new entropy term, named *tautomeric entropy*, that results from the explicit inclusion of tautomerism in the simulations and how it can be computed together with the occupational entropy. Simulations using tautomerism were done for hen egg white lysozyme (HEWL) using a simple set of tautomers at dihedral energy minima. A very good overall prediction of pK_a values was obtained, presumably the best in the literature for HEWL, using a high value for the dielectric constant assigned to the protein region, ϵ_p. The explicit inclusion of water molecules treated under the extended tautomer formalism further improved the prediction, in contrast with previous works using rigid water molecules. In all calculations performed, the region with $\epsilon_p \approx 20$ is shown the to be the optimal one. Some aspects of the somewhat controversial issue of the "proper" ϵ_p value are also discussed.

1. Introduction

Solution pH is a major determinant of protein structure and function.[1,2] The effect of pH arises through changes in the protonation equilibrium of titrable sites, essentially of electrostatic nature,[3] and has long prompted the development of theoretical models aimed at understanding and predicting those properties[4−11] Recent protonation modeling studies in proteins have included the prediction of pK_a values,[9−25] the calculation of pH-dependent stability curves,[15,17,26] and studies of the coupling of protonation with redox processes.[27−31]

Simulations of the global protonation equilibrium of proteins, usually based on continuum electrostatic (CE) methods,[32−34] commonly treat protons in two different ways. The first approach assigns an average charge to the protonated form of the protonatable sites, either by using a single unitary charge[9,17] or by distributing the proton partial charges over the site atoms.[21,28,30] However, protonatable sites often have alternative proton positions, and therefore, even by using an average proton distribution, this approach is unable to describe the specific interactions that may be established by sites among themselves and with other hydrogen-bond donor/acceptor groups. The second approach tries to deal with this aspect by selecting for each site one of the proton positions as the one appropriated to the protonated form.[14,16,19,25,30,35] However, the relative populations of the alternative proton positions for one site will in general depend on the state of all other sites, so the a priori choice of a particular proton location for a site, even if reasonable for a particular global proton configuration, may become invalid under different conditions. Furthermore, besides this more obvious energetic aspect, the existence of alternative proton positions has entropic consequences (see below) which cannot also be accounted for by the usual approaches. A proper consideration of proton isomerism seems therefore necessary.

The first inclusion of proton isomerism in CE-based simulations of the protonation equilibrium was the treatment of His tautomerism in myoglobin.[13] Among the protonatable sites usually found in proteins, His is the only one that displays true proton tautomerism in solution. However, the usual CE approach uses a rigid protein structure (but see below), and therefore, even sites not displaying proton tautomerism in solution have to be treated as tautomeric[36] because interconversion through rotation around single bonds is not possible in a rigid molecule. This means that virtually all sites become tautomeric: e.g., carboxyl sites can bind the proton to any of its two oxygen atoms (eventually with variable geometries), while amino sites can lose any of its three protons. An inclusion of this type of generalized proton tautomerism in protonation equilibrium calculations has been recently done by Alexov and Gunner,[23] who observed an improved prediction of pK_a values for hen egg white lysozyme. The general purpose of the present work is to investigate some theoretical and computational issues related with the inclusion of proton tautomerism; four major aims can be identified, which are discussed below.

The treatment of proton tautomerism within the CE framework can be done by splitting each site into several pseudosites, which are then treated using the usual methodology;[13] site−site interactions must be set in such a way that forbidden protonation states (e.g., doubly deprotonated His sites) never

* Corresponding author. Address: António M. Baptista, Instituto de Tecnologia Química e Biológica, Av. da República, EAN, ITQB II, Piso 6, Apartado 127, 2781-901 Oeiras, Portugal. Tel.: 351−214469613. Fax: 351−214411277. E-mail: baptista@itqb.unl.pt.

10.1021/jp002763e CCC: $20.00
Published on Web 12/07/2000

occur. The aim is to transform the tautomeric system into an equivalent nontautomeric one, which could then be treated in the conventional way. Besides its theoretical interest, this approach has the advantage that the more widespread computational tools for treating the nontautomeric case can be used to study tautomeric systems as well. However, the general transformation between the two systems was until now not described in detail. The first aim of this work is to provide a rigorous link between the tautomeric and nontautomeric systems. In particular, the choice of a proper reference state in the pseudosite approach is discussed. The question of the proper pK_a values of tautomeric model compounds is also discussed, which may be relevant to other approaches not using the pseudosite formalism.[23,26,38,39]

The inclusion of proton tautomerism in protonation equilibrium calculations allows for some reorganization of the protein to the protonation changes themselves. Even though a more satisfactory treatment of reorganization certainly requires the consideration of more extensive local[18,20,23,39] or global[12,15,19,25,40−43] conformational aspects, proton tautomerism leads to a significant improvement of computed pK_a values, as shown by Alexov and Gunner[23] and the present work (see below). Although the emphasis was previously put on energetic aspects,[23] the thermodynamic reasons for this improvement are both energetic and entropic. The second aim of this work is to discuss the origin and practical importance of these entropic consequences of proton tautomerism. In particular, it is shown that entropic corrections should be included from the start in model pK_a values, rather than using an approximate a posteriori correction.[23]

The approach used to deal with the proton tautomerism of protonatable sites can be extended to treat other groups whose protons are not uniquely determined from the known structures, such as alcohol and (free) thiol groups, and even bound water molecules;[23] the occurrence of their charged forms just needs to be forbidden. Previous work has shown that the inclusion of rigid water molecules seems to overestimate the associated pK_a shifts,[14] suggesting that some form of partially rigid treatment would be the best solution. A more detailed study has been done by Gibas and Subramanian,[22] who concluded that the explicit inclusion of rigid water molecules buried in the protein had little effect in pK_a prediction, while the inclusion of additional ones proved detrimental; the eventual change of hydrogen-bond networks upon (de)protonation was not treated. Given these results, we may expect that the use of the proton tautomerism treatment for the water molecules could result in better pK_a predictions. However, previous work following this approach has not examined this issue in detail, since only a few water molecules have been included and no comparison has been done with calculations without explicit water.[23] The third aim of this work is to investigate how extensive the explicit inclusion of water molecules can be, using the proton tautomerism approach.

Presently, CE methods seem to provide the only feasible route to address the global protonation equilibrium of a protein because the large number of protonation free energies that are required can be efficiently computed as sums of individual and pairwise terms, whose calculation is much faster than that allowed by other methods. However, the use of a single protein conformer leads to some problems because the protein conformation cannot reorganize in response to the protonation changes. The more theoretically sound solution to this problem within the context of CE methods is perhaps to restrict CE calculations to rigid structures and combine them with molecular dynamics (MD) algorithms, obtaining a method for constant-pH MD simulations;[40] unfortunately, this approach is still prohibitive to studying the global protonation equilibrium over a wide pH range. A limited set of conformations can also be selected from NMR measurements[19] or MD simulations[12,15,41,43] and used for direct averaging; however, besides being in limited number, the conformations are in principle more characteristic of the original protonation state(s) and may loose their adequacy over the pH range. A method for the weighting of conformers, based on the pH-induced shift of the relative conformer probabilities, has also been proposed;[25] the previous limitations also apply here. A self-consistent minimization of the structure using average protonations has also been suggested,[42] but this is clearly an approximation for the conformation−protonation coupling occurring in solution. Another solution is to allow for some local flexibility of the protein, usually by selecting alternative site conformers and including molecular mechanics (MM) energy terms directly in the protonation energies;[18,20,23,39] unfortunately, this approach is always necessarily approximate because the MM terms have to be included without affecting the linear pairwise nature of the CE energies. In face of the theoretical and/or practical drawbacks of all these methods, an alternative (and computationally faster) solution to the conformational problem is the use of a high value (∼20) for the dielectric constant of the protein region, ϵ_p, which several workers found to compensate for the rigidity of the structure.[17,21,44] The reasons for this improvement are not totally clear, but they are usually assumed to be related with the fact that the flexibility of the protein (especially at the surface) allows for more extensive reorientation of dipoles and charges of the protein and also for solvent penetration; roughly speaking, both phenomena correspond to an increase of the dielectric constant in that region (at least when responding to an external field[45]). In their treatment of proton tautomerism, Alexov and Gunner[23] suggest that proton conformational isomerism (which in their case is more extensive than the purely tautomeric treatment considered here) partially plays the role of a high dielectric constant, thus explaining the quality of the pK_a values predicted using $\epsilon_p = 4$. The fourth aim of this work is to investigate the effect of tautomerism on the dielectric dependence of pK_a predictions. It is shown that although tautomerism clearly improves predictions, the dielectric profile remains unchanged. In particular, the optimal ϵ_p remains around 20. The general question of the "proper" ϵ_p to be used in CE-based simulations is also discussed.

The protein used here to study the aforementioned aspects was hen egg white lysozyme (HEWL). Due to the existence of accurate experimental data,[46,47] this protein has over the years become a standard test system for pK_a prediction methodologies.[6,9,11,15−23,43,48,49]

2. Theory

2.1. Equivalent Nontautomeric System. The proton occupation state of a protein with N protonatable sites can be represented as a vector $\mathbf{n} = (n_1, n_2, ..., n_N)$, where $n_i = 0$ or 1 indicates that site i is respectively empty or occupied. The probability of the protein being in state $\mathbf{n}$ is given by the general expression for (semi-)open systems,[50] which in this case becomes

$$p(\mathbf{n}) = \frac{\exp[-2.3n\mathrm{pH} - \Delta G(\mathbf{n})/kT]}{\sum_{\mathbf{n}'} \exp[-2.3n'\mathrm{pH} - \Delta G(\mathbf{n}')/kT]} \qquad (1)$$

where $\Delta G(\mathbf{n})$ is the standard free energy of the protonation

reaction $\mathbf{0} \rightarrow \mathbf{n}$ and $n = \sum_i n_i$ is the total number of bound protons; the summation in the denominator extends over all 2^N protonation states and is equal to the semi-grand canonical partition function of the system, relative to the fully deprotonated state. When tautomerism does not exist, $\Delta G(\mathbf{n})$ can be pairwise decomposed through CE quantities[12,14,17,26,49]

$$\Delta G(\mathbf{n}) = -2.3kT\sum_i n_i \mathrm{p}K_i^{\mathrm{int}} + \frac{1}{2}\sum_i\sum_j (n_i n_j + n_i z_j^\circ + n_j z_i^\circ) W_{ij} \quad (2)$$

where 2.3 stands for ln 10, and z_i° is the charge (in protonic units) of site i in the deprotonated form. The intrinsic pK_a of site i, pK_i^{int}, is its pK_a when all the others are kept neutral, and W_{ij} is the CE "interaction" between the occupied sites i and j (e.g., see refs. 14 and 26 for the exact definition). We note that $W_{ij} = W_{ji}$ and that for simplicity of notation in the summations, we have defined $W_{ii} = 0$. It is possible to include non-CE (MM) energy terms into $\Delta G(\mathbf{n})$,[18,20,23,39,42] but in the present work, this inclusion is not necessary because of the nature of the alternative proton positions being considered (see Experimental Methods section).

When proton tautomerism exists, the occupation state $\mathbf{n}$ no longer gives a complete description of the protonation state, since one or both states of each n_i may have alternative forms. The protonation state can be fully represented as a vector $\mathbf{x} = (x_1, x_2, ..., x_N)$, where x_i can take as many values as the total number of protonated and deprotonated states (e.g., three for a His site).[26,38] As stated in the Introduction, one of our tasks is to find a thermodynamically equivalent system of nontautomeric pseudosites which has the same statistical mechanical properties as those of the original system of tautomeric sites. If we designate the occupation state of this equivalent nontautomeric system as $\mathbf{m}$ (defined similarly to $\mathbf{n}$ above), we need to define a transformation $\mathbf{x} \rightarrow \mathbf{m}$ and the $\Delta G(\mathbf{m})$ values in such a way that $p(\mathbf{x}) = p(\mathbf{m})$.

One way of doing the transformation $\mathbf{x} \rightarrow \mathbf{m}$ is to decompose each tautomeric site into several nontautomeric pseudosites, more exactly one per tautomer.[13] For example, a His site will be represented as two pseudosites: one that can be charged or in the neutral $N^{\epsilon 2}$-deprotonated form, and another that can be charged or in the neutral $N^{\delta 1}$-deprotonated form. Since this equivalent system is nontautomeric, we can use eq1 and write the probabilities of the $\mathbf{m}$ states as

$$p(\mathbf{m}) = \frac{\exp[-2.3m\mathrm{pH} - \Delta G(\mathbf{m})/kT]}{\sum_{\mathbf{m}'} \exp[-2.3m'\mathrm{pH} - \Delta G(\mathbf{m}')/kT]} \quad (3)$$

where $\Delta G(\mathbf{m})$ has to be expressed in terms of CE quantities. This has to be done in a way that preserves the energetics of the original tautomeric system while satisfying three important conditions. First (i), we must choose a well-defined reference state. The state with all pseudosites neutral, used as the reference state in nontautomeric calculations,[9,14] is not uniquely defined for the protonatable sites usually found in proteins. On the contrary, the nontautomeric form of the usual sites is the charged one (carboxyl, amino, phenyl, and thiol groups all have several neutral forms but a single charged form, at least if only dihedral energy minima are considered), which suggests that the charged form is a more convenient reference state. Hence, assuming we restrict our attention to sites whose tautomeric forms are neutral, each tautomer should be modeled as a pseudosite whose reference state is the charged form. Second (ii), a pseudosite can interact with no more than one charged pseudosite from another site and with none from its own site. Third (iii), sites cannot have more than one pseudosite in the neutral state (e.g., a His site cannot be doubly deprotonated). In other words, the $\mathbf{m}$ states not realized by the transformation $\mathbf{x} \rightarrow \mathbf{m}$ must have $p(\mathbf{m}) = 0$. If we manage to define $\mathbf{x} \rightarrow \mathbf{m}$ to satisfy these three conditions while preserving the original CE energetics of the system, we obtain an equivalent system which necessarily satisfies $p(\mathbf{x}) = p(\mathbf{m})$.

Condition i demands that, for each pseudosite r, we use as a reference its pK_a when all other sites are charged (i.e., when $m_{s\neq r} = z_s^\circ + 1$), which we represent as pK_r^*. The corresponding protonation reaction, whose free energy is $-2.3kT\mathrm{p}K_r^*$, can be described by the two-step process

$$\{m_r = 0, \text{ all } m_{s\neq r} = z_s^\circ + 1\} \rightarrow \{\text{all } m_s = 0\} \rightarrow \{m_r = 1, \text{ all } m_{s\neq r} = z_s^\circ + 1\} \quad (4)$$

The first step is a deprotonation and the second a protonation. Since the equivalent system is nontautomeric, we can use eq2 (substituting $\mathbf{m}$ for $\mathbf{n}$) to compute the free energies of these steps, whose sum gives

$$-2.3kT\mathrm{p}K_r^* = -2.3kT\mathrm{p}K_r^{\mathrm{int}} + \sum_s (2z_s^\circ + 1) W_{rs} \quad (5)$$

Since $2z_r^\circ + 1 = \pm 1$, depending on whether the pseudosite is cationic or anionic, this equation can be regarded as obtained by adding/subtracting the "interactions" of r (i.e., the W_{rs} terms) with the cationic/anionic pseudosites, which is what one would expect intuitively. By inserting eq 5 into eq 2 (substituting $\mathbf{m}$ for $\mathbf{n}$) and rearranging terms, we get

$$\Delta G(\mathbf{m}) = -2.3kT\sum_r m_r \mathrm{p}K_r^* + \frac{1}{2}\sum_r\sum_s (m_r m_s + m_r z_s^* + m_s z_r^*) W_{rs} \quad (6)$$

where we have defined $z_r^* = -z_r^\circ - 1$. The introduction of these charges can be regarded as the "reversal" of the charge types of the pseudosites (cationic $\leftrightarrow$ anionic); W_{rs} is invariant with respect to such reversal. In this way, we obtained an expression for $\Delta G(\mathbf{m})$ whose reference state is the charged protein. The expression was intentionally written in a form identical to eq 2, as in the nontautomeric case.

As for condition ii, we first note that the term in parentheses in the double summation in eq 6 can be written as

$$(m_r + z_r^*)(m_s + z_s^*) - z_r^* z_s^* \quad (7)$$

The last constant term has no effect on the probabilities, and therefore, we can concentrate our attention on the first product. But when a pseudosite r is charged, we have $(m_r + z_r^*) = 0$, and thus, this product vanishes. Therefore, a charged pseudosite never effectively interacts (*i.e.*, never affects the probabilities), and condition ii is always satisfied.

With respect to condition iii, we note that when two pseudosites r and s corresponding to the same site i ($r, s \in i$) are simultaneously neutral, the first product in eq 7 is equal to one. Hence, in order to avoid the double neutral form, we just need to define a very high W_{rs} value because $\Delta G(\mathbf{m})$ then becomes very high also, and $p(\mathbf{m}) \rightarrow 0$. Obviously, to avoid numerical problems caused by the high W_{rs} values, we should truly remove the constant $z_r^* z_s^*$ terms from the calculation of the

probabilities; this is already a common procedure in nontautomeric calculations.[9,31,48]

Since eq 6 is formally identical to eq 2, the treatment of proton tautomerism can in fact be done using the computational tools developed for the nontautomeric case, as assumed in previous works.[13] As shown above, such an approach requires (a) the decomposition of each tautomeric site into nontautomeric pseudosites (one per tautomer), (b) the redefinition of the pseudosites types ($z_r^\circ \rightarrow z_r^*$) and values ($pK_r^{int} \rightarrow pK_r^*$), and (c) the imposition of very high (infinite) W_{rs} values between pseudosites corresponding to the same site. We note also that the transformation $\mathbf{x} \rightarrow \mathbf{m}$ just described is equally valid if $\mathbf{x}$ denotes more extensive configurational changes; in that case, the original pK_r^{int} and W_{rs} will typically include MM energy terms.[20,38,39]

Finally, we note that the occupation state n_i of a tautomeric site i can be easily obtained from the pseudosite occupations by using the fact that no more than one of its pseudosites can be neutral. If site i (and its pseudosites) is anionic, multiple occupied pseudosites are impossible, so that

$$n_i = \sum_{r \in i} m_r \tag{8}$$

If the site is cationic, multiple empty pseudosites are impossible, so

$$1 - n_i = \sum_{r \in i} (1 - m_r) \tag{9}$$

or

$$n_i = 1 - \tau_i + \sum_{r \in i} m_r \tag{10}$$

where τ_i is the number of pseudosites (and tautomers) of site i.

2.2. Model Compound pK_a Values. The pK^{int} values are usually computed from the known pK_a values of model compounds (pK^{mod}) by means of a thermodynamic cycle and CE calculations.[7,12,14,17,26] However, these pK^{mod} values cannot be used for the pseudosites defined above, since their usual values refer to a tautomeric (i.e., pseudosite averaged) site.

From the existence of a single charged form and several neutral ones, it follows from simple chemical equilibrium arguments that the global (averaged) K_i^{mod} of a tautomeric site i is given by

$$(K_i^{mod})^{2z_i^\circ + 1} = \sum_{r \in i} (K_r^{mod})^{2z_i^\circ + 1} \tag{11}$$

where K_r^{mod} refers to pseudosite r; we note that $2z_i^\circ + 1 = \pm 1$. When there are energetic differences between the alternative neutral forms (as in His), we need to know their relative populations in order to compute the individual K_r^{mod} values; in this case, some kind of experimental data is necessary, as used for the study of His tautomerism in myoglobin.[13] When all neutral forms are energetically equivalent, all K_r^{mod} values are necessarily equal and we get

$$pK_r^{mod} = pK_i^{mod} + (2z_i^\circ + 1) \ln \tau_i \tag{12}$$

In this case the difference between pK_i^{mod} and pK_r^{mod} is purely entropic.

When the proper pK^{mod} values for the pseudosites are used from the start, the a posteriori approximate entropy correction suggested by Alexov and Gunner[23] becomes unnecessary.

2.3. Entropy Decomposition. The entropy of a system of tautomeric protonatable sites can be written as[50]

$$S = -k \sum_{\mathbf{x}} \sum_{\Gamma} p(\mathbf{x},\Gamma) \ln p(\mathbf{x},\Gamma) \tag{13}$$

where Γ represents the configurational state of the system (protein and solvent coordinates and momenta), assumed to take discrete values for the simplicity of notation; the order of the summations is significant, since in general Γ depends on $\mathbf{x}$. To analyze the different entropy contributions, it is, however, convenient to express the protonation vector $\mathbf{x}$ in an alternative way. For this purpose, we introduce an additional tautomeric state vector $\mathbf{t} = (t_1, t_2, ..., t_N)$, where $t_i = t_i(n_i)$ can take as many values as the total number of alternative tautomers of the site i when in occupation state n_i (e.g., a His site will have one t_i value for $n_i = 1$ and two values for $n_i = 0$). Hence, we have $\mathbf{x} = (\mathbf{n},\mathbf{t})$, with $\mathbf{t} = \mathbf{t}(\mathbf{n})$. Analogously to the nontautomeric case,[31] the entropy of the system can now be written as $S = S^{occ} + \langle S(\mathbf{n}) \rangle$, where

$$S^{occ} = -k \sum_{\mathbf{n}} p(\mathbf{n}) \ln p(\mathbf{n}) \tag{14}$$

is the occupational entropy of the system,[31] due to the occupied/empty fluctuations of the sites ($n_i = 1/0$), and $\langle S(\mathbf{n}) \rangle$ is the average of

$$S(\mathbf{n}) = -k \sum_{\mathbf{t}} \sum_{\Gamma} p(\mathbf{t},\Gamma|\mathbf{n}) \ln p(\mathbf{t},\Gamma|\mathbf{n}) \tag{15}$$

the conditional entropy of the system in occupation state n. In the nontautomeric case $S(\mathbf{n})$ is only due to fluctuations at the configurational level;[31] in the present case, it reflects also the fluctuations at the tautomeric level and can be decomposed as $S(\mathbf{n}) = S^{taut}(\mathbf{n}) + S^{conf}(\mathbf{n})$, where

$$S^{taut}(\mathbf{n}) = -k \sum_{\mathbf{t}} p(\mathbf{t}|\mathbf{n}) \ln p(\mathbf{t}|\mathbf{n}) \tag{16}$$

is the conditional entropy due to fluctuations at the tautomeric level and $S^{conf}(\mathbf{n}) = \langle S(\mathbf{n},\mathbf{t}) \rangle_{\mathbf{n}}$ is the conditional, tautomer-averaged value of

$$S(\mathbf{n},\mathbf{t}) = -k \sum_{\Gamma} p(\Gamma|\mathbf{n},\mathbf{t}) \ln p(\Gamma|\mathbf{n},\mathbf{t}) \tag{17}$$

the latter being the conditional entropy of the system in state $(\mathbf{n},\mathbf{t})$, due only to fluctuations at the configurational level. If we define the tautomeric entropy $S^{taut} = \langle S^{taut}(\mathbf{n}) \rangle$ and the configurational entropy $S^{conf} = \langle S^{conf}(\mathbf{n}) \rangle$, we can finally write

$$S = S^{occ} + S^{taut} + S^{conf} \tag{18}$$

Thus, when compared to the nontautomeric case,[31] the inclusion of proton tautomerism introduces a new entropic contribution S^{taut}, which is qualitatively different from the other two.

As previously discussed,[31] the occupational entropy has a clear thermodynamic meaning, since it reflects the effect of deconstraining an extensive parameter of the system: the number of bound protons. Although part of this entropy comes from the existence of alternative states with the same number of protons, the physical reason these alternative states can occur is that we have "turned on" the labile character of the protons that turns the system into an open one. Hence, S^{occ} is essentially of thermodynamic origin and does not result from an arbitrary separation of the degrees of freedom, as is often the case for

entropy terms (e.g., the translational, rotational, and vibrational entropies of an ideal polyatomic gas).

As for S^{taut}, which reflects the tautomeric fluctuations displayed by fixed occupation states, it can be regarded as an appropriation of part of the true configurational entropy because for all sites except His, the alternative tautomers should be interconvertible through rotation around single bonds (unless nonbonded constraints exist). This means that by regarding all sites as tautomeric, we exclude at least part of that rotational freedom from S^{conf}. Hence, as with most entropy decompositions, the split of the nonoccupational entropy into S^{taut} and S^{conf} results essentially from a convenient separation of degrees of freedom. Of course, this separation is motivated by the fact that the standard one-conformer CE approach is probably unable to account for fluctuations of this magnitude (see Introduction). The inclusion of tautomerism in CE-based simulations of protonation equilibrium is thus a way of allowing for some of the (total) conformational reorganization in response to protonation changes. The contribution of S^{taut} to any free energy change cannot be obtained by using a single, even if well-chosen, tautomeric state for each site in rigid CE-based simulations. Even if we choose the most likely set of tautomeric sites so that the energy terms become reasonably estimated, the entropic effect of tautomeric fluctuations will be necessarily ignored and will be lacking in the computed free energies. This role of explicitly added degrees of freedom as a source of further generalized fluctuations and of increased ability to reorganize, which is a general statistical mechanical effect, has already been discussed in the context of protonation equilibrium in proteins.[31,40]

It is of interest to compute the nonconfigurational entropic contributions for the protonation reactions of individual sites, as done previously for S^{occ}.[31] Although it does not seem possible to uncouple S^{occ} and S^{taut} in terms of computable quantities, their sum is easily obtained. We first note that

$$S^{occ} + S^{taut} = -k\sum_{\mathbf{n}}\sum_{\mathbf{t}} p(\mathbf{n},\mathbf{t}) \ln p(\mathbf{n},\mathbf{t}) = -k\sum_{\mathbf{x}} p(\mathbf{x}) \ln p(\mathbf{x}) \quad (19)$$

Since the transformation $\mathbf{x} \rightarrow \mathbf{m}$ preserves the probabilities of the protonation states, it follows that

$$S^{occ} + S^{taut} = -k\sum_{\mathbf{m}} p(\mathbf{m}) \ln p(\mathbf{m}) \quad (20)$$

In other words, the sum of the occupational and tautomeric entropies of the tautomeric system is equal to the occupational entropy on the nontautomeric equivalent system. From this result we can derive the joint contribution of S^{occ} and S^{taut} to the standard entropy change of the reaction of occupation of a site i ($n_i = 0 \rightarrow 1$); following a derivation analogous to the one in the appendix of ref 31, we obtain

$$T\Delta S_i^{occ} + T\Delta S_i^{taut} = \frac{\mathrm{cov}[n_i, \Delta G(\mathbf{m}) + 2.3kT m\mathrm{pH}]}{\mathrm{var}(n_i)} + kT \ln \frac{\langle n_i \rangle}{1 - \langle n_i \rangle} \quad (21)$$

Furthermore, the free energy of this reaction is given by

$$\Delta G_i = -2.3kT\mathrm{pH} - kT \ln \frac{\langle n_i \rangle}{1 - \langle n_i \rangle} \quad (22)$$

Therefore, we can compute ΔG_i and its contributions $T\Delta S_i^{occ} + T\Delta S_i^{taut}$ and $\Delta H_i - T\Delta S_i^{conf}$ from the binding statistics of the system. The first of these contributions is particularly interesting because its $T\Delta S_i^{taut}$ part measures a contribution which, as noted above, cannot possibly be accounted for in a nontautomeric CE-based simulation. We note also that when analyzing the contributions for the midpoint pK_a of the site, pK_i^{half}, only the first term needs to be considered in the two previous equations.

The present entropy decomposition is equally valid if **x** denotes more extensive configurational changes,[20,38,39] with the vector **t** representing the explicit configurational degrees of freedom. As above, Γ will stand for the hidden configurational states.

3. Computational Methods

3.1. Model Structure. The structure used as a starting point for the calculations was the triclinic form of HEWL from Hodsdon et al.,[51] deposited as entry 1LZT in the Protein Data Bank.[52] Selected crystallographic water molecules were used in some of the calculations (see Results and Discussion), chosen according to their individual solvent accessibilities. This is obtained for each water molecule by removing all other water molecules and computing individual accessibilities using the program ASC;[53,54] the water molecule number 31 in the PDB file was discarded, since it is only 0.07 Å away from water molecule number 14. To avoid periodic boundary artifacts, we performed the selection of water molecules using the water molecules of all 27 unitary cells forming a cube, of which the cell containing the protein is the center; the translation operations were done using the program Swiss-PdbViewer.[55]

The addition of hydrogen atoms to the structure was done in a way that generates all the protons to be used in the tautomeric calculations; following Alexov and Gunner,[23] we treated the Ser and Thr residues and water molecules (when included) as protonatable sites. Hydrogen atoms were first automatically added using GROMOS 87,[56] including explicit polar and aromatic hydrogens,[57] assuming the protonated form for all sites; crystallographic water molecules were not considered at this stage. When conformational freedom of the protons exists (all sites except His and Arg), this results in their placement at the positions corresponding to a minimum value of the dihedral conformational energy. To account for the alternative tautomer geometries, "excess" protons were added as follows. In the case of carboxyl sites (Asp, Glu, and C-terminus), a symmetric copy of the automatically placed proton was added to the deprotonated carboxyl oxygen. In the case of Tyr, an additional proton was added to the phenyl oxygen at the other minimum of the torsion potential (at 180° from the dihedral of the automatically placed proton). In the case of sites with alcohol groups (Ser and Thr), two additional protons were added to the hydroxyl oxygen at the other two minima of the torsion potential (at ±120° from the dihedral of the automatically placed proton); this procedure can also be used for free Cys sites, which do not exist in HEWL. Finally, each water molecule (when these were included; see Results and Discussion) was taken as the center of a randomly oriented tetrahedron and one proton placed at each vertex. All these operations were done using GROMOS 87 geometries.[56,57]

The above procedure leads to protons at conformational minima and randomly oriented water molecules. This is a simpler set of positions than the ones previously used, obtained by optimization of hydrogen bonds and minimization of electrostatic energy.[23] Although the present procedure may seem to generate less realistic positions, it must be noted that in the

case of nonwater sites, the physically favored proton positions cannot in general show extreme deviations from dihedral minima, given the resultant energy cost; a similar situation seems to hold with respect to heavy atoms.[58−60] In addition, crystal structures (including crystallographic water molecules) may show non-negligible deviations from the protein structure in solution (which could explain the better CE-based pK_a predictions using an average, relaxed structure[43]), so many of the proton positions derived from details of the structure could be meaningless; thus, the procedure adopted here produces protonation−conformation couplings that are probably more robust toward structural errors than methods using a local optimization. Finally, the use of dihedral energy minima avoids the use of MM energy terms (see below), making the method much easier to implement.

3.2. Protonation free energies. As mentioned in the Theory section, non-CE terms can be included in the protonation free energies.[18,20,23,39,42] However, with the present methodology, all proton positions of a given site correspond to equivalent conformational minima (see above) and thus have the same conformational energy. Furthermore, in the GROMOS 87 force field,[56,57] used to derive all parameters (see below), these protons have no Lennard−Jones interaction energy. Hence, the nonelectrostatic part of the GROMOS 87 energy differs between protonation states only through differences in the Lennard−Jones parameters of the atoms to which the titrable protons are bound; this difference was ignored in this work, and only one set of parameters was used for the different states (the charged set for carboxyl, amino, and His sites; the neutral set for Tyr sites). Under this assumption, the protonation free energies $\Delta G(\mathbf{n})$ could be computed using only CE terms.[9,14] Furthermore, since the Lennard−Jones parameters are used to derive the atomic radii for the CE calculations (see below), this assumption also ensures that the dielectric boundary of the protein becomes independent of the protonation states.

The protons added to the original structure (see above) enable considering several tautomers for each original site. In general, each cationic/anionic site that is truly titrable can have one of its proton locations empty/occupied, which results in two tautomers for His, Tyr, and carboxyl sites and three for amino sites. For Ser and Thr sites, we defined three tautomers, each having a single proton at one of the three alternative positions. For water sites, we defined six tautomers, corresponding to pairs of the four proton positions. Each of the tautomers thus defined was treated as an independent pseudosite in the subsequent calculations. Arg sites were considered to be charged and nontitrable.

The protonation free energy terms were computed using the program MEAD (version 1.1.8),[12,61] which uses a finite difference method to solve the linear Poisson−Boltzmann equation. Each pK_r^* value was computed using an individual run with all other sites in the charged state. Additional calculations were done to obtain all cross-terms W_{rs}, which were subsequently changed for pseudosite pairs in the same site (see Theory section). The calculations were done with an ionic strength of 0.1 M, a temperature of 300 K, a molecular surface[62] defined with a solvent probe radius of 1.4 Å, and a Stern (ion exclusion) layer of 2.0 Å. A two-step focusing procedure[63] was used, with consecutive grid spacings of 1.0 and 0.25 Å. The dielectric constant of the solvent was 80 and that of the protein (ϵ_p) varied between 2 and 80.

The pK_r^{mod} values required for the calculations were computed as follows. Sites which are actually not titrable (Ser, Thr, and water) were assigned to p$K_r^{mod} = -100$ and subsequently treated as cationic; this guarantees that they remain neutral. For the truly titrable sites except His, the pK_r^{mod} value was obtained from eq 12, using the pK_i^{mod} values previously used for nontautomeric calculations.[30] In the case of His, where the two tautomers are energetically different, we used the pK_r^{mod} values of 7.0 and 6.6 for the $N^{\delta 1}$ and $N^{\epsilon 2}$ tautomers, respectively, as adapted by Bashford et al.[13] from NMR data;[64] this corresponds to p$K_i^{mod} = 6.45$, using eq 11.

Atomic radii were taken to be half of the distance corresponding to the minimum Lennard−Jones energy between like-atom pairs in the GROMOS 87 force field.[56,57] Almost all atomic partial charges were taken from this same force field; the exception was deprotonated Tyr, whose atomic partial charges were the ones previously used.[30] For Ser, Thr, and water, the atomic partial charges in the protonated forms are irrelevant, although MEAD needs a charge difference to identify the site type; in these cases, the charge of the actual proton(s) plus a small formal charge was uniformly distributed over the alternative proton positions. For the purpose of comparison (see Results), some calculations were done using average charged forms (see Introduction). In this case, the partial charges for truly titrable sites were the ones previously reported,[30] while for Ser and Thr sites the GROMOS 87 values were modified by assigning −0.549 to the hydroxyl oxygen and 0.133 to each of the three alternative proton positions.

3.3. Monte Carlo Titrations. The sampling of protonation states was performed using a Monte Carlo (MC) method, as done in previous works.[14,31,48,65] As seen in the Theory section, this sampling can be done in terms of the states **m** of the nontautomeric pseudosites. Consequently, there is in principle no need to develop new algorithms and computational programs for the tautomeric case, since the nontautomeric versions can be used. Nevertheless, the use of such algorithms, though theoretically valid, can result in some practical problems. In the first place, many changes of state can occur only through the flip of several pseudosites because the change between two neutral tautomers of a given site requires that the currently occupied pseudosite becomes empty and the one corresponding to the new tautomer becomes occupied. If an MC algorithm with double flips for strongly coupled sites[48] is used, the pseudosites of the same tautomeric site will be selected for double MC moves (because their W_{rs} values are very high; see Theory section), and therefore, the tautomerizations should in principle be possible. However, besides not guaranteeing a good sampling, this results in a very high number of coupled pseudopairs, which may render the method computationally too demanding for large proteins. Furthermore, even though the use of pseudosite double moves makes the tautomerizations possible, the occurrence of all variants of double moves involving two true sites will be possible only if we introduce higher order moves: e.g., two sites with three neutral tautomers each would require multiple moves involving all six pseudosites. These higher-order moves may not be absolutely necessary, but their lack will probably lead to a poor sampling, as does the lack of double moves in nontautomeric sampling.[48] Finally, the tautomeric nature of the pseudosites leads to many impossible multiple moves that, although rejected by the use of high W_{rs} vales (see Theory), consume a large part of the computation time, especially if water molecules are treated as tautomeric sites.

The simplest solution to this problem is perhaps to restate it in terms of **x** states (defined at the beginning of the Theory section). We start by splitting eq6 in terms of tautomeric sites

$$\Delta G(\mathbf{m}) = \sum_i [-2.3kT \sum_{r\in i} m_r \mathrm{p}K_r^*] + \frac{1}{2}\sum_i \sum_j \left[\sum_{r\in i}\sum_{s\in j}(m_r m_s + m_r z_s^* + m_s z_r^*) W_{rs}\right] \quad (23)$$

After rearrangement of the terms and since each **x** state corresponds to a particular **m**, we can switch to the **x** states and write $\Delta G(\mathbf{x})$ as

$$\Delta G(\mathbf{x}) = \sum_i g_i(x_i) + \frac{1}{2}\sum_i \sum_{j\neq i} g_{ij}(x_i, x_j) \quad (24)$$

where we have defined

$$g_i(x_i) = -2.3kT \sum_{r\in i} m_r \mathrm{p}K_r^* + \frac{1}{2} g_{ii}(x_i, x_i) \quad (25)$$

$$g_{ij}(x_i, x_j) = \sum_{r\in i}\sum_{s\in j}(m_r m_s + m_r z_s^* + m_s z_r^*) W_{rs} \quad (26)$$

The previous calculation of the quantities $g_i(x_i)$ and $g_{ij}(x_i,x_j)$ makes possible the sampling in terms of **x** states, which results in a much more efficient sampling and a decrease of about one order of magnitude in the computation times. In this approach, one may select only valid **x** states, which makes unnecessary the artificial use of very high W_{rs} values between pseudosites $r, s \in i$. In fact, the term $^1/_2 g_{ii}(x_i,x_i)$ in $g_i(x_i)$ should be discarded because, strictly speaking, it does not contribute to the true protonation free energy $\Delta G(\mathbf{x})$.

The MC sampling of **x** states was done based on eq 24, using the Metropolis scheme.[66,67] Trial moves consisted of random choice of x_i states. Single site moves were used for all sites and double site moves were used for strongly coupled sites,[48] defined here as the ones with at least one $|W_{rs}| > 2$pK units ($r \in i$, $s \in j$). As before,[31] instead of a random selection of individual and paired sites, a sequential scheme was used, which also leads to a proper Markov chain.[68] The calculation of pK^{half} values, directly given by protonation averages, was done using 4×10^4 MC steps, where one MC step is defined as a full cycle of trial moves over the list of individual and paired sites. Unless stated otherwise, tautomer populations and entropies were computed using 10^6 MC steps. In all cases, simulations were done for $-5 \leq$ pH ≤ 25 using intervals of 0.1.

Although the results presented here were all obtained with the above procedure using sampling of **x** states, the direct use of the MC program previously developed for the nontautomeric case (which includes double moves)[31] yielded the same results for all cases compared. This may not be true for proteins with stronger site−site couplings than HEWL because of the sampling reasons discussed above. When many water molecules are introduced (see Computational Results and Discussion), a very high number of coupled pseudosite pairs results, and the calculations using the original MC algorithm become prohibitively slow.

4. Results and Discussion

4.1. Prediction of pK^{half} Values. As discussed in the Introduction, the results of CE-based simulations of protonation equilibrium depend on the value used for ϵ_p, the dielectric constant of the protein interior.[17,21,44] Hence, the analysis of the effect of the inclusion of proton tautomerism on the computed pK^{half} values is here done considering this dependence. A more detailed discussion of the dielectric aspects is given before the Conclusions.

A comparison was done between the tautomeric treatment proposed here and the two usual ways of modeling titrating protons (see Introduction): as an average entity collapsed on the remaining atoms and as a single tautomer. Two single-tautomer runs, hereafter designated as single1 and single2, were done for all sites using the first and the second tautomers in the program list, respectively. Since this numbering is based on the atom names, there is no reason to expect any physically meaningful correlation of these choices with the sites of the protein, and therefore, this is essentially equivalent to using two random selections of tautomers. No water molecules were used in these runs, and the tautomeric run included "tautomers" for sites with free alcohol groups (Ser and Thr; see Computational Methods). Figure 1 shows the root-mean-square deviation (RMSD) between the calculated and experimental pK^{half}s as a function of ϵ_p for the different representations of the protons; since experimental ranges, rather than values, are reported in some cases, each RMSD was computed using the nearest value of the experimental range. It is clear from the figure that the use of a single proton position can lead to very different prediction accuracies, depending on the particular choice of tautomers. Also, this approach can lead to better or worse results than the average approach, depending again on the chosen tautomers. The predictions using the tautomeric approach are systematically better than those of the other approaches, giving an overall improvement of about 0.1 pK units over single1 (the best nontautomeric result) in the region of lower RMSD. Thus, besides the theoretical reasons for its use, the tautomeric methodology proposed here does actually compute better pK^{half} values than the usual, nontautomeric approaches.

Table 1 shows the experimental pK^{half} ranges for all sites and the corresponding values calculated using $\epsilon_p = 20$, with the tautomeric and nontautomeric runs. It can be seen that for many sites, the best calculated pK^{half} is not the tautomeric one; nevertheless, the best overall prediction is the tautomeric one, as was already evident from Figure 1. The sites giving worse pK^{half} values in the nontautomeric runs, namely, the N-terminus, Glu-35 (at the active site) and Asp-66, do not show much improvement when tautomers are used, showing that some important factors are still missing in the modeling of the system.

As discussed in Computational Methods, the tautomeric treatment can be used for truly protonatable sites and also for sites with free alcohol groups (Ser and Thr). Figure 2 shows the global RMSD obtained with and without alcohol tautomers (the latter using the original proton placed by GROMOS; see Experimental Methods). The RMSD curves for most individual sites (not shown) are indeed very similar. The main exception is Asp-48, whose RMSD increases by about 0.8 in the $\epsilon_p = 20$ region when alcohol tautomers are excluded; this site has a strong direct interaction with Ser-50 (max $W_{rs} = 0.65$ pK units with $\epsilon_p = 20$), which is responsible for this effect (see below). Other sites showing non-negligible but smaller differences are Asp-87, Tyr-20, N-terminus, Asp-52, and Asp-66; all of them have also significant interactions with at least one alcohol group. Hence, the improvement of pK^{half} predictions obtained for HEWL using the tautomeric treatment comes in general from the truly protonatable sites, although the consideration of alcohols may be crucial for some sites.

Comparison of the present results with previous CE-based calculations on HEWL is not straightforward for several reasons. In the first place, some calculations were done for the tetragonal crystal form[15,17,43] and others for triclinic forms different from that used here.[17,19,21,22] Second, the "preparation" of the structure (addition of hydrogens, etc.) that precedes the actual calculations

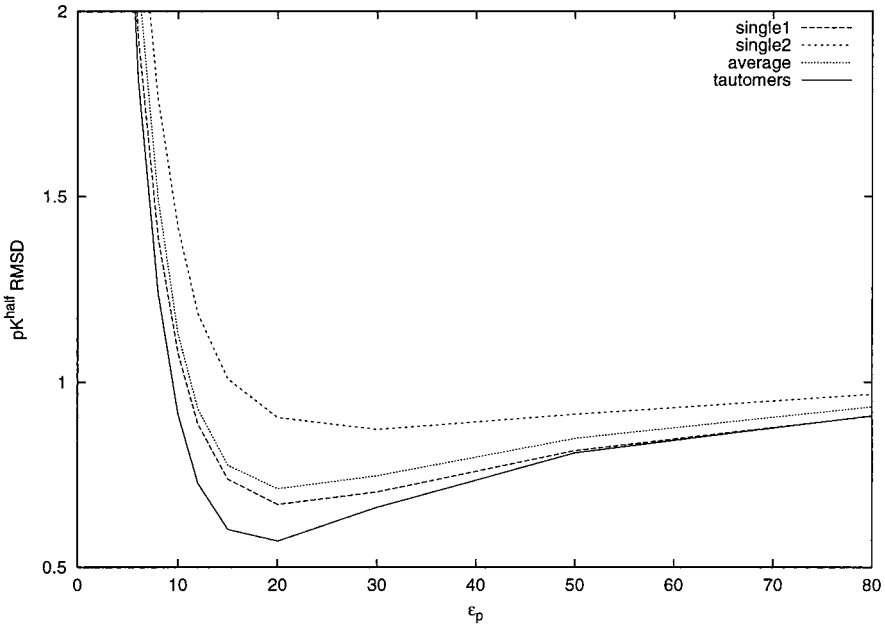

Figure 1. RMSD between experimental and calculated pK^{half} values with and without tautomers. See text for further details.

TABLE 1: Effect of Tautomerism on Calculated pK^{half} Values

			calculated ($\epsilon_p = 20$)[a]		
site	experimental[b]	tautomers	single 1	single 2	average
N-terminus	7.8−8.0	7.02	7.09	6.48	6.86
Lys-1	10.7−10.9	10.51	10.40	10.45	10.45
Glu-7	2.60−3.10	2.93	2.80	3.00	2.92
Lys-13	10.4−10.6	11.66	11.66	11.73	11.67
His-15	5.29−5.43	5.52	5.73	4.65	5.00
Asp-18	2.58−2.74	3.18	3.08	3.32	3.09
Tyr-20	10.3	9.93	10.22	10.10	10.10
Tyr-23	9.8	9.79	9.83	9.72	9.83
Lys-33	10.5−10.7	10.04	9.97	10.21	9.98
Glu-35	6.1−6.3	4.52	4.69	4.22	4.36
Asp-48	1.2−2.0	2.10	3.25	1.74	2.73
Asp-52	3.60−3.76	3.40	2.36	3.52	3.51
Tyr-53	12.1	12.27	12.29	11.52	11.83
Asp-66	0.4−1.4	2.41	2.50	4.27	3.30
Asp-87	1.92−2.22	1.92	2.19	2.51	2.32
Lys-96	10.7−10.9	11.29	11.40	11.19	11.18
Lys-97	10.2−10.4	11.22	11.06	11.33	11.20
Asp-101	4.02−4.16	3.81	3.78	3.83	4.12
Lys-116	10.3−10.5	10.34	10.38	10.28	10.32
Asp-119	3.11−3.29	3.50	3.59	3.44	3.52
C-terminus	2.63−2.87	2.72	2.94	2.44	2.63
RMSD	−	0.57	0.67	0.90	0.71

[a] Errors lower than 0.005 pK_a units. [b] Experimental values for Asp, Glu, His, and C-terminus are from ref. 47 and the remaining ones from ref 46.

is done using protocols that obviously vary. Furthermore, besides the ϵ_p value, the CE methodology depends on several parameters, namely, the sets of charges and radii;[13,21,69] even the choice of ionic strength can be a problem because the structure determination and the (often multiple) pK_a measurements do not usually refer to the same conditions. Finally, some calculations do not report the full set of pK^{half} values for which experimental values exist, so a full comparison is not possible. Nevertheless, Table 2 shows the total RMSD of pK^{half} values from several studies where a reasonable comparison is possible; the so-called "null model", corresponding to using pK^{half} = pK^{mod} for all sites,[17] is also shown. The RMSD values are computed in the same manner as that in Table 1, i.e., with respect to the experimental range, using all available pK^{half} values. The studies of Yang et al. and Antosiewicz et al. are standard calculations using rigid crystal structures. The study of Alexov and Gunner is of particular interest, since it includes proton conformational isomerism, similarly to our treatment of tautomers. The calculations of Beroza and Case also include flexibility, using an extended conformation for each site as an alternative to the crystal one. The calculation of Vlijmen et al. uses an average structure from an MD simulation, which the authors found to give the best results among different averaging alternatives. The perhaps most immediate feature in Table 2 is that calculations with $\epsilon_p = 20$ give consistently better results than those with $\epsilon_p = 4$; previous studies comparing different ϵ_p values also pointed to a high optimal value.[17,21,26] Moreover, while the calculations with high ϵ_p roughly halve the error of the null model, the calculations with low ϵ_p rank similar to this much simpler model. Thus, the use of a high ϵ_p value seems to consistently improve the results, independent of the methodological details; this will be discussed in more detail below. These results show also that our simple inclusion of conformational effects using alternative tautomers at dihedral minima (see Computational Methods), together with the use of a high ϵ_p, is at least as good as the other conformational treatments in the table, namely, those of Beroza and Case, Alexov and Gunner, and Vlijmen et al. The comparison with the studies listed in Table 2 is probably even more favorable to our methodology than suggested by the RMSDs shown, since several of them do not report the pK^{half} values of some problematic sites (see table footnotes). Nevertheless, and despite the apparently better performance of the method proposed here, the different methodologies used in these studies make definite conclusion obviously impossible. This problem is discussed below in more detail.

4.2. Tautomer Populations. As observed by Alexov and Gunner,[23] the populations of alternative tautomers change with pH, due to the new reciprocal stabilizations originated by the titration of particular sites. Our results are very similar to theirs, in particular the large variations observed within the Asp-66 and Asp-87 "clusters".[23] We show in Figure 3 some tautomer populations for those two clusters and another one involving Asp-48, already mentioned above. The general reasons for the fluctuations in tautomer populations have been previously

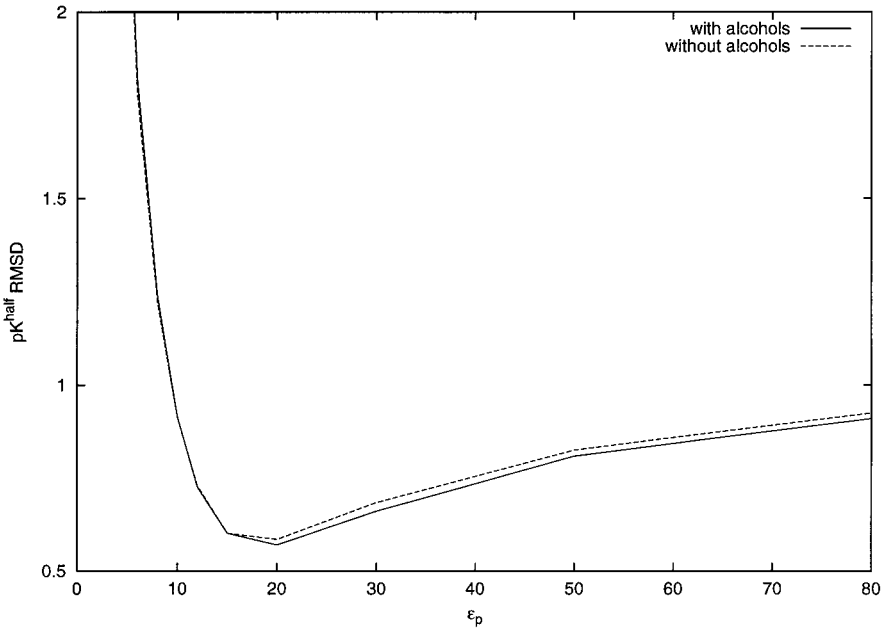

Figure 2. RMSD between experimental and calculated pK^{half} values with and without the inclusion of alcohols as tautomeric sites.

TABLE 2: pK^{half} RMSDs from HEWL Calculations

calculation	ϵ_p	RMSD
Null model	–	1.18
Yang etal.[15]	4	1.12[a]
Antosiewicz et al.[17]	20	0.65[b]
Beroza and Case[20]	4	1.39
Alexov and Gunner[23]	4	>1.40[c]
Vlijmen et al[43]	20	0.61
this work with $\epsilon_p = 20$	20	0.57

[a] No pK^{half} reported for Tyr sites. [b] No pK^{half} reported for N-terminus and Asp-119. [c] pK^{half} for Tyr-53 is reported only as >16.

addressed by Alexov and Gunner,[23] and we will focus instead on the differences observed with respect to their work, essentially due to our different choice of tautomers.

The Asp-66 cluster involves several sites which interact strongly with this site and among themselves, namely, Tyr-53, Thr-89, and, with weaker interactions, Ser-60 and Thr-51; in particular, a very high interaction couples the states of Asp-66 and Tyr-53 strongly together. The changes we obtain for Tyr-53 and Thr-69 are not as marked as the ones observed by Alexov and Gunner.[23] In the case of Tyr-53, this is possibly due to the fact that the tautomers used by these authors are minimized with respect to local (electrostatic) energy while ours simply use the usual dihedral minima (see Computational Methods). In the case of Thr-69 and in addition to this effect, we use three tautomers instead of the two used by these authors; the curves for two of them follow each other very closely, and roughly speaking, together they replace one of their tautomers. Our prediction of the pK^{half} of Asp-66 (Table 1) is clearly worse than the value (1.5) obtained by these authors, suggesting that their choice of tautomers is better in this case.

The Asp-87 cluster involves this site, Thr-89, and His-15; weak interactions exist also among His-15, Tyr-20, and several Lys sites. In this cluster, we again observe changes smaller than those of Alexov and Gunner;[23] in particular, we do not observe a second inversion of two Thr-89 tautomers. The reason here is probably again the fact that we use three instead of two tautomers, using different criteria. Even if we sum the populations of the two tautomers with lower populations at high pH, the inversion is still not observed. Our prediction of the pK^{half} of Asp-87 (Table 1) is clearly better than the value (0.5) obtained by these authors, suggesting that our choice of tautomers is better in this case.

The Asp-48 cluster (not discussed by Alexov and Gunner[23]) can be actually regarded as part of the Asp-66 cluster, with which it interacts through Tyr-53. The strongest interaction is between Asp-48 and Ser-50. Here we observe an inversion of the populations of two of the tautomers of Ser-50. Our prediction of the pK^{half} of Asp-48 (Table 1) is only slightly worse than the value (1.3) obtained by these authors, suggesting that the choice of tautomers is not determinant in this case.

Overall, the tautomer populations observed seem to suggest that the simple definition of alternative tautomers used in this work can generally capture the proper energetics of the system.

4.3. Entropic Contributions. As discussed in the Theory section, the explicit treatment of tautomers naturally takes into account a new entropic term, S^{taut}, which is necessarily absent from other approaches, even if the latter somehow use tautomer-averaged energies. Figure 4 shows the nonconfigurational entropy contributions for the pK^{half} values of all sites for $\epsilon_p = 20$. This corresponds to ΔS_i^{occ} for the nontautomeric runs (dashed and dotted lines) and to $\Delta S_i^{occ} + \Delta S_i^{taut}$ for the tautomeric one (thin solid line); all are computed using eq 21, since the absence of tautomerism can be treated as a particular case where a single tautomer exists.

The nontautomeric curves are very similar, with most differences smaller than 0.1 pK units (except for Asp-48 and Asp-52), showing that the alternative nontautomeric treatments are very similar in terms of entropic consequences. At first sight, the inclusion of tautomerism seems to have drastic effects, given the large positive and negative $T\Delta S_i^{occ} + T\Delta S_i^{taut}$ values for some sites. However, it must be noted that the very protonation of the site corresponds to a change between a neutral state with several alternative tautomers and a charged state with a single form, which necessarily results in a significant contribution to ΔS_i^{taut}. This contribution from the site itself is somewhat the protein counterpart of the (exact) correction done to its pK_i^{mod} value (see Theory). This self-contribution to ΔS_i^{taut} cannot be formally separated from it because, as usual, the pairwise nature of the (free) energy of the system is lost due to the mixing

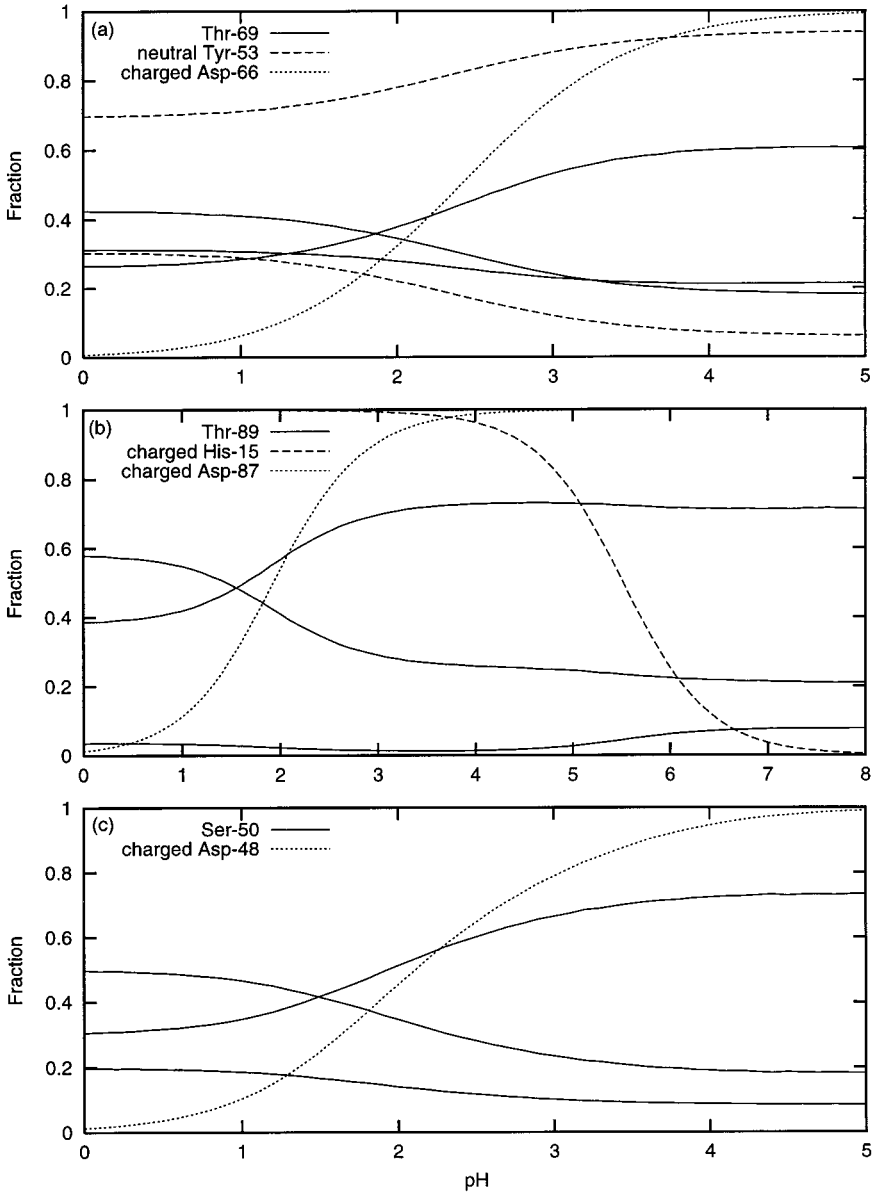

Figure 3. Selected tautomer populations for some strongly interacting sites, with $\epsilon_p = 20$. (a) Asp-66, Tyr-53, and Thr-69. (b) Asp-87, Thr-89, and His-15. (c) Asp-48 and Ser-50.

character of the entropy. We can, however, define a "self-tautomeric entropy" of a site i with occupation n_i as

$$S_i^*(n_i) = -k\sum_{t_i} p(t_i|n_i) \ln p(t_i|n_i), \qquad (27)$$

which, strictly speaking, corresponds to ignoring the tautomeric nature of all other sites. The contribution to ΔS_i^{taut} we are seeking can then be approximately written as $\Delta S_i^* = S^*(1) - S^*(0)$, with one of the terms (the one referring to the charged state) being zero. Although this quantity is not a proper component of ΔS_i^{taut}, it is an approximate measure of the physical effect we want to capture. It reaches its maximum absolute value when all tautomers are equally likely, namely, $|\Delta S_i^*| = k \ln \tau_i$, which corresponds to about 0.3 and 0.5 pK units for sites with two and three alternative tautomers, respectively. In the case of interest here (the contribution to pK_i^{half} values), we must use the probabilities observed for the tautomers when $\langle n_i \rangle = {}^1/_2$. The "corrected" tautomeric entropy contributions thus obtained, without the self-tautomeric term ΔS_i^*, are also shown in Figure 4 (thick solid line). This curve is indeed much closer to the nontautomeric ones, as expected. However, some significant deviations are observed, the larger ones being for sites Tyr-20 (~0.1 pK units) and Asp-52 and Asp-66 (~0.2 pK units for both). Asp-66 was already referred to above with respect to changes in tautomer populations, being the core of the cluster with strongest interactions. It interacts very strongly with Tyr-53 and more weakly with Thr-69, Ser-60, and Thr-51; thus, it is not surprising that its protonation affects significantly the tautomer populations of the other sites, as already seen in Figure 3a. The protonation of Asp-52 affects also the populations of the neutral Tyr-53 and also those of Ser-50. The protonation of Tyr-20 has a large effect on the tautomer populations of Ser-100 and, to a less extent, on those of the neutral His-15.

These results show that the protonation of a site can in general lead to significant tautomeric rearrangements of other sites in the protein. This means that even if we could somehow devise a way of estimating the midpoint tautomer populations of a site i using a nontautomeric treatment (certainly not an easy task),

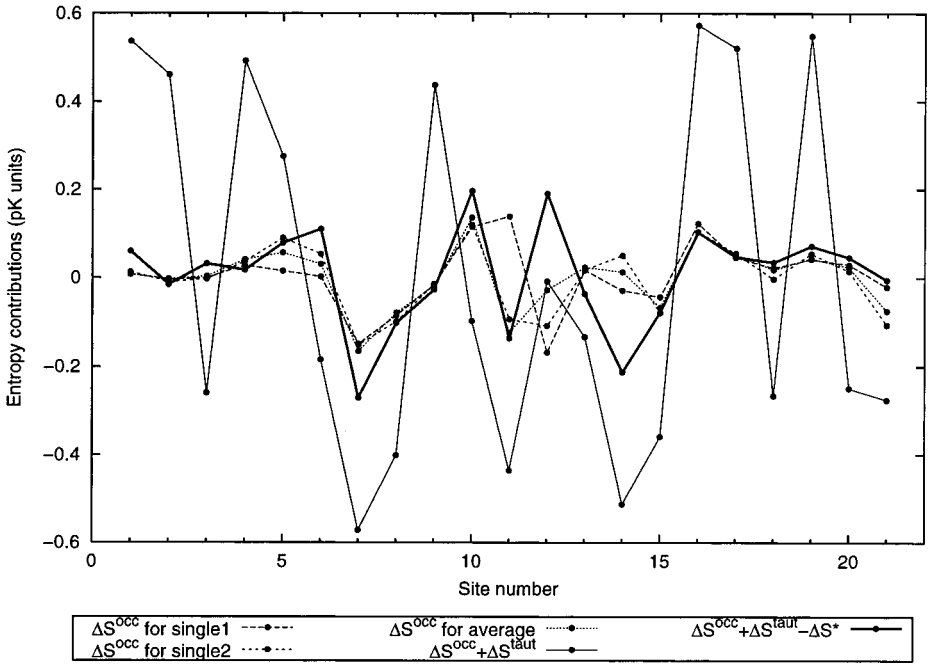

Figure 4. Entropy contributions for pK^{half} values, for tautomeric (solid lines) and nontautomeric (nonsolid lines) calculations, using $\epsilon_p = 20$. Site numbers refer to the order in the tables. See text for further details.

that would not be enough to compute the ΔS_i^{taut} missing in the calculated pK_i^{half} value; we would still have errors of at least ± 0.2 pK units. Thus, we conclude that ΔS_i^{taut} cannot be easily obtained except through an explicit consideration of tautomerism, as done here.

4.4. Explicit Inclusion of Water Molecules. As mentioned in the Introduction, water molecules can be explicitly included in the calculations as tautomeric sites, analogous to what was done for alcohol groups in the runs discussed above. The water molecules thus modeled are essentially orientable permanent dipoles, the extent and change of their orientation being limited only by the number of tautomers used (see Computational Methods). In this way, the solvent is modeled as a combination of orientable water molecules and a dielectric continuum. This can be seen as akin of the PDLD/S method of Warshel and co-workers,[10] where the solvent is modeled as a combination of Langevin dipoles and a dielectric continuum, although our "dipoles" are in smaller number and with a much more limited orientation.

Several sets of calculations were done, each including water molecules selected with an individual relative solvent accessibility below a given cutoff value, whose ϵ_p-dependent pK^{half} RMSD profiles are shown in Figure 5. Only the region of lower RMSD values is shown, the overall curve shape being similar in all cases where the full ϵ_p range was scanned; due to computational reasons, only some ϵ_p runs were done for the higher cutoffs. Up to a cutoff of 0.3, the curves always remain below that corresponding to no explicit waters (cutoff = 0.0); i.e., the inclusion of water molecules with a relative accessibility below 0.3 always leads to improved pK^{half} predictions, regardless of the ϵ_p value. Above that cutoff value, this feature is lost, but the calculations showing the lowest RMSD are still the ones with $\epsilon_p = 20$. In all cases, the RMSD at $\epsilon_p = 20$ is smaller than that obtained without explicit water, having its smaller value for an accessibility cutoff of 0.5. These results are in marked contrast with those of Gibas and Subramanian,[22] who observed that the inclusion of increasing numbers of waters molecules, modeled as rigid, led to a progressively worse global pK^{half} prediction in HEWL. Since the main difference between the two studies is the treatment of waters, the origin of the present improvement is certainly the use of orientable (tautomeric) water molecules in our calculations. The explicit orientable water molecules can in principle give rise to a more realistic dielectric response of the solvent, including charge-dependent and nonlinear effects,[70−74] which cannot possibly be modeled using a solvent dielectric constant.

Table 3 shows the effect of the inclusion of explicit water molecules in the individual pK^{half} predictions, using $\epsilon_p = 20$. Despite the overall improvement with respect to the calculations without water molecules, there is no consistent trend caused by this inclusion; i.e., the pK^{half} values of some sites improve, while others become worse. In particular, of the problematic sites N-terminus, Glu-35, and Asp-66 (see above), only Glu-35 shows a significant and consistent improvement with increasing cutoff. These results show that although the inclusion of tautomeric water molecules is in general advantageous, it may lead in some cases to a worse local model of the solvent than the use of a continuum. Clearly, a more extensive study of the selection criteria would be of interest.

Given the nonhomogeneous nature of proteins, different behaviors are expected for the several water molecules included in the simulations. In particular, it is interesting to examine the degree of orientation of the water molecules and to see if it differs much from the rigid orientations used in the study of Gibas and Subramanian.[22] The instantaneous orientation of a water molecule can be measured by its relative dipole vector, $\boldsymbol{\mu}/\mu$, which is a unit vector. Thus, a convenient measure of the degree of orientation of a water molecule is the norm of its average relative dipole, $||\langle \boldsymbol{\mu}/\mu \rangle||$, which takes values between 0 (totally disordered) and 1 (rigid orientation). It is also interesting to examine the orientability of the water molecules, i.e., their ability to adapt to particular instantaneous states and thus contribute to relaxation. This orientability can be measured as the fluctuation (standard deviation) of the relative dipole. However, this fluctuation is equal to $(1 - ||\langle \boldsymbol{\mu}/\mu \rangle||^2)^{1/2}$, and therefore, the two measures are actually equivalent; i.e., a high (low) degree of orientation implies a low (high) orientability. There is thus no need to directly examine the fluctuations. Figure

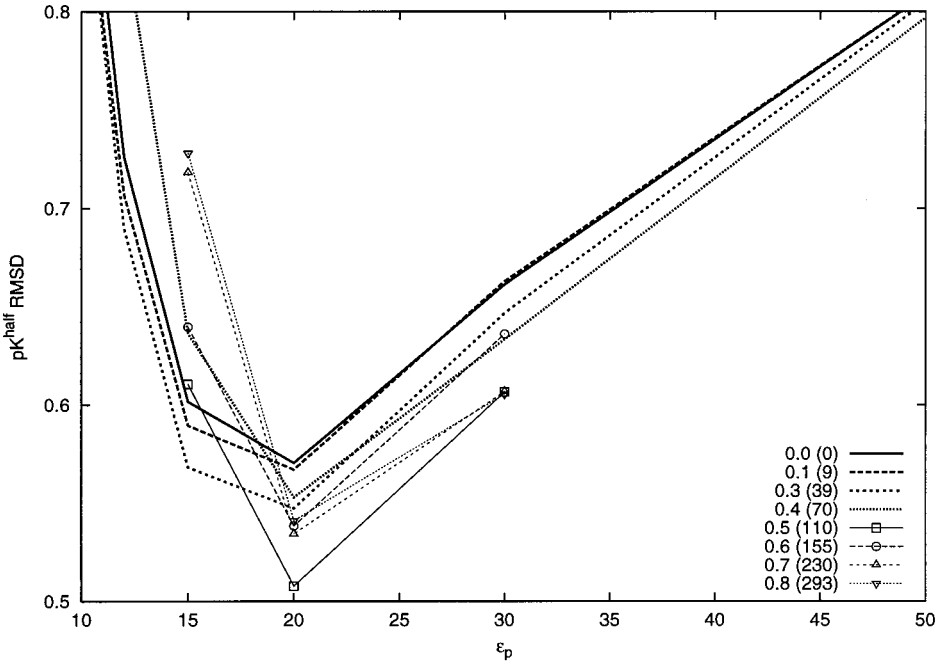

Figure 5. RMSD between experimental and calculated pK^{half} values using different cutoffs of relative solvent accessibility for the inclusion of water molecules. The cutoff of 0.0 corresponds to the exclusion of all water molecules. The number of selected water molecules is shown in parentheses. See text for further details.

TABLE 3: Effect of Water Inclusion on Calculated pK^{half} Values[a]

		calculated ($\epsilon_p = 20$)			
site	experimental	no water	acc <0.3	acc <0.5	acc <0.7
N-terminus	7.8−8.0	7.02	7.14	6.99	6.87
Lys-1	10.7−10.9	10.51	10.47	10.37	10.20
Glu-7	2.60−3.10	2.93	2.92	2.90	2.84
Lys-13	10.4−10.6	11.66	11.71	11.63	11.56
His-15	5.29−5.43	5.52	5.52	5.42	5.21
Asp-18	2.58−2.74	3.18	3.07	2.89	3.26
Tyr-20	10.3	9.93	9.95	10.43	10.42
Tyr-23	9.8	9.79	9.82	9.84	10.12
Lys-33	10.5−10.7	10.04	9.99	9.84	9.82
Glu-35	6.1−6.3	4.52	4.63	5.07	5.16
Asp-48	1.2−2.0	2.10	2.09	2.01	2.04
Asp-52	3.60−3.76	3.40	3.44	3.35	3.00
Tyr-53	12.1	12.27	12.28	12.54	12.70
Asp-66	0.4−1.4	2.41	2.40	2.43	2.27
Asp-87	1.92−2.22	1.92	2.02	1.98	1.88
Lys-96	10.7−10.9	11.29	11.27	11.24	11.15
Lys-97	10.2−10.4	11.22	11.18	11.12	10.89
Asp-101	4.02−4.16	3.81	3.84	3.92	4.06
Lys-116	10.3−10.5	10.34	10.32	10.33	9.85
Asp-119	3.11−3.29	3.50	3.52	3.59	3.69
C-terminus	2.63−2.87	2.72	2.74	2.79	2.39
RMSD	−	0.57	0.55	0.51	0.53

[a] See notes in Table 1.

6 shows calculated $||\langle\boldsymbol{\mu}/\mu\rangle||$ values for selected water molecules using $\epsilon_p = 20$ and an accessibility cutoff of 0.5. Before analyzing these results, it must be noted that the computed $||\langle\boldsymbol{\mu}/\mu\rangle||$ values are in general lower than the ones which could in principle be obtained using totally orientable water molecules (i.e., with a virtually infinite number of tautomers). This is due to the fact that in terms of the calculations, the configuration of a particular water molecule is limited to the six tautomers used to model it (see Computational Methods). These six configurations depend on the orientation of the tetrahedron used to generate them (see Computational Methods) or, equivalently, on the orientation of the corresponding octahedron formed by the six resulting dipole directions. Since this orientation is randomly generated (see Computational Methods), it may not include some of the most electrostatically favorable directions. The worse situation to model corresponds to have an overwhelmingly favorable (i.e., essentially unique) dipole direction pointing to a face of the octahedron; in the simulation, this direction will be "substituted" by the closest ones, namely, the three pointing to the vertexes of that face, which will be equally populated, leading to $||\langle\boldsymbol{\mu}/\mu\rangle|| = 1/\sqrt{3} \approx 0.58$. A similar, though less severe, underestimation of $||\langle\boldsymbol{\mu}/\mu\rangle||$ may occur for other cases, unless the random orientation happens to satisfy approximately the most favorable configuration(s). Hence, despite the expected advantages of using non-"optimized", randomly oriented water molecules (see Computational Methods), this approach may result in the underestimation of some of the larger $||\langle\boldsymbol{\mu}/\mu\rangle||$ values.

The water molecules featured in Figure 6 illustrate the wide variety of behaviors with respect to orientation. Some of them display large changes with pH, due to strong interactions with titrable groups. Water molecule 99 (HOH-99) is in a totally rigid orientation at low pH, due to its strong interaction with the charged N-terminus (max $W_{rs} = 1.32$ pK units); as the latter becomes neutral, the interaction between them and the resulting average orientation become lower. HOH-4 has a large interaction with Asp-18 (max $W_{rs} = 0.62$ pK units), which leads also to a large shift in $||\langle\boldsymbol{\mu}/\mu\rangle||$ as this site titrates. HOH-64 has a large interaction with Tyr-53 (max $W_{rs} = 0.78$ pK units) and a weaker one with Asp-66 (max $W_{rs} = 0.18$ pK units); thus, its $||\langle\boldsymbol{\mu}/\mu\rangle||$ curve shows two shifts in the corresponding pH titrating regions of these two residues. The pH-dependent behavior of these water molecules is obviously impossible to describe by modeling them as rigid and can only be captured by using alternative orientations. On the other hand, other water molecules have an essentially constant $||\langle\boldsymbol{\mu}/\mu\rangle||$ over the whole pH range. HOH-36 has a strong interaction with Arg-5, which, being treated as nontitrable (see Computational Methods), is always charged. HOH-1 has a low interaction with Tyr-53 (max $W_{rs} = 0.10$ pK units) and a moderate one with Asp-66 (max $W_{rs} = 0.30$ pK units), remaining nearly constant. The $||\langle\boldsymbol{\mu}/\mu\rangle||$ of HOH-62 is also essentially constant, with no significant interactions with

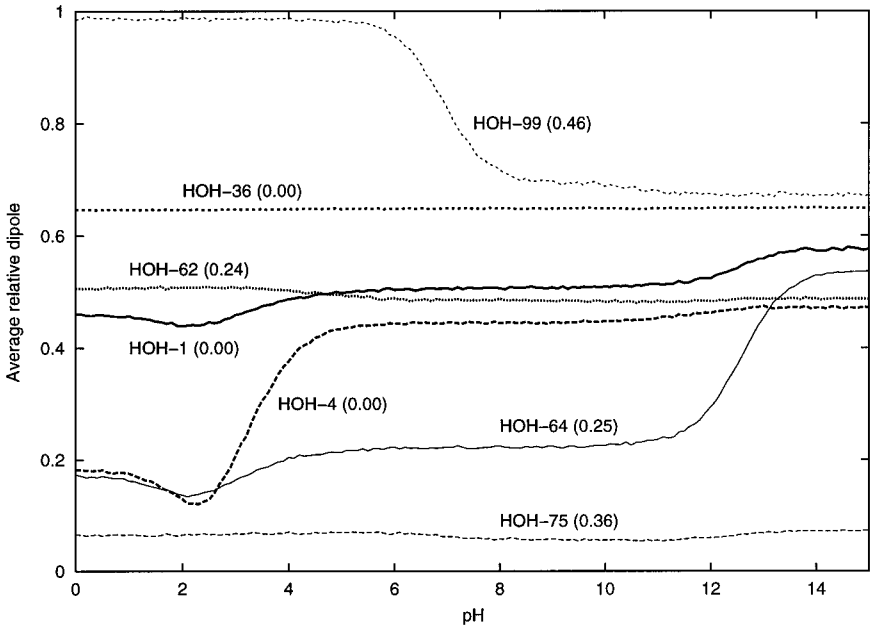

Figure 6. Average relative dipole, $||\langle\boldsymbol{\mu}/\mu\rangle||$, of selected water molecules as a function of pH, using $\epsilon_p = 20$. "Sequence" numbers are from the PDB file (only water molecules from the original unitary cell are shown; see Computational Methods). Relative individual accessibilities are shown in parentheses. The number of MC steps was 4×10^5.

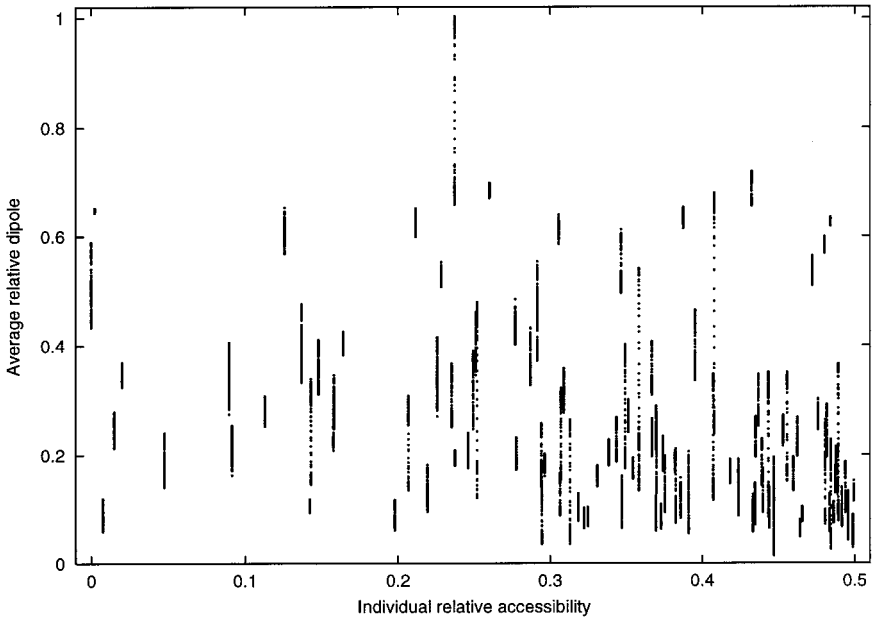

Figure 7. Average relative dipole, $||\langle\boldsymbol{\mu}/\mu\rangle||$, as a function of the individual relative solvent accessibility, for all water molecules, using $\epsilon_p = 20$. Points for all pH values between 0 and 15 are shown, the columns of points corresponding to the spanning of this range for individual water molecules.

truly titrable sites. Hence, HOH-36, HOH-1, and HOH-62 have a roughly constant high $||\langle\boldsymbol{\mu}/\mu\rangle||$ and could probably be reasonably treated using rigid water molecules. Finally, HOH-75 behaves essentially as a bulk water molecule, with no favored orientation. There is thus a large variety of behaviors with respect to the orientation of water molecules, ranging from the ones that are essentially rigid to those indistinguishable from bulk ones.

An interesting feature of Figure 6 is that there seems to be no obvious correlation between $||\langle\boldsymbol{\mu}/\mu\rangle||$ and the individual relative accessibility of the water molecule. This is confirmed by Figure 7, which shows the relation between these two properties for all water molecules and for all pH values in the range from 0 to 15. Although some tendency for small $||\langle\boldsymbol{\mu}/\mu\rangle||$ values is observed for the more exposed waters, a similar tendency is not observed for the more buried ones. Indeed, as shown in Figure 8, the distribution of $||\langle\boldsymbol{\mu}/\mu\rangle||$ values is roughly uniform at an accessibility cutoff of 0.2, the increasing of the cutoff leading to a gradual increase of the population of low values. Hence, there is a tendency for exposed water molecules to behave as bulk, although buried ones do not tend to behave as rigidly oriented ones.

4.5. "Proper" ϵ_p Value. As shown by this and previous works,[17,21,26] CE-based simulations of protonation equilibrium depend markedly on the value of ϵ_p. This naturally prompts the questions why this is so and what the "proper" ϵ_p value is a recurring theme in the literature.[8,75−78] In our opinion, the essential point is the fact that the meaning and value of a

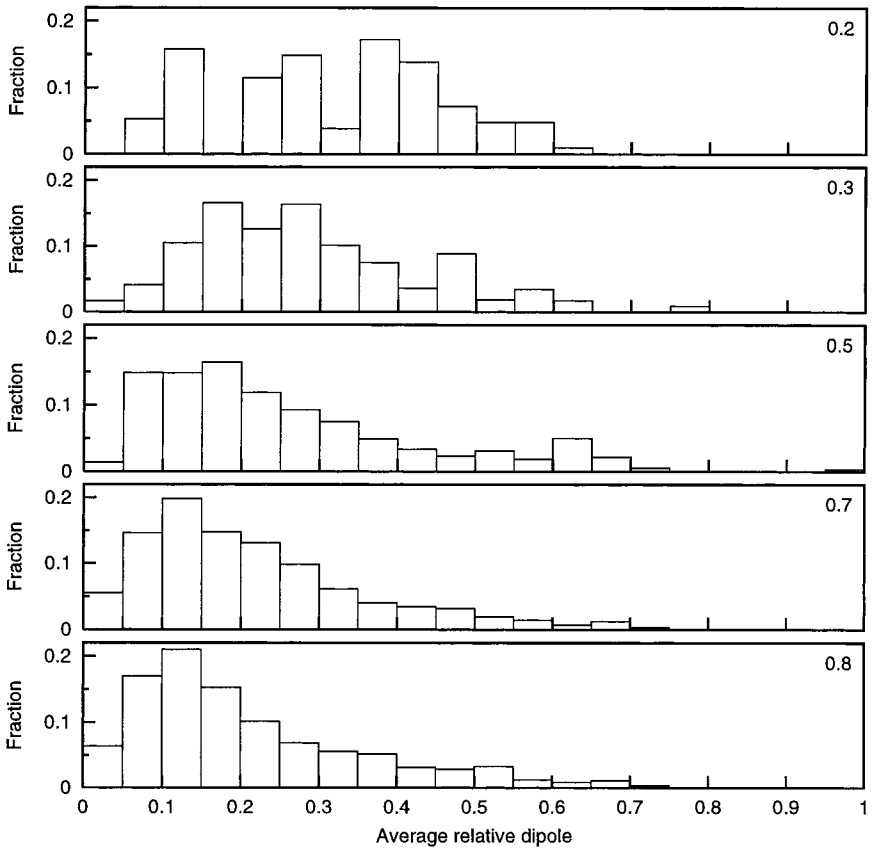

Figure 8. Populations of water molecules in terms of their average relative dipole, $||\langle \boldsymbol{\mu}/\mu \rangle||$, for different accessibility cutoffs (shown in upper right corner), using $\epsilon_p = 20$. All pH values between 0 and 15 are considered, so that the histograms reflect the global distribution within that pH range.

dielectric constant depend on its definition, as repeatedly pointed out by Warshel and co-workers,[8,24,77,79] Many dielectric constants are used in theoretical studies of protein systems, from the "effective" ones obtained from direct application of some equation (Coulomb's law,[80,81] Tanford−Kirkwood model,[5,82] or generalized Poisson−Boltzmann equation[34,83]) to the ones based on a theory that establishes a clear link with the polarizability of the underlying atomic system (the Kirkwood−Fröhlich theory of dielectrics[75,78,84,85] or an atomic-based linear response formalism[86]). In all cases, the use of a dielectric constant is an attempt to reflect aspects of a system that are not treated in an explicit way. In the case of CE models, ϵ_p (which is essentially equivalent to ϵ_{in} in the PDLD/S method[10]) accounts for electronic polarization and also for the structural aspects not explicitly included in the model of both the protein and the surface water. These structural aspects are not only the displacements associated with the reorganization to charge changes,[24,79] but also the fluctuations occurring in the system with a particular charge state (including the neutral one). The parameter ϵ_p tries to capture these aspects in an global way, over all sites and pH range of interest. Hence, in the absence of more detailed models, ϵ_p may be considered essentially an adjustable parameter whose optimal value must be found by comparison of predictions with experimental data.

Previous studies where several ϵ_p were tested[17,21,26] have all shown that a high ϵ_p of about 20 or more gives the best pK^{half} predictions for several proteins. This high value can be interpreted (approximately, in terms of the response to external fields) as due to the fact that the flexible protein structure can display a significant response to changes in electric field by adapting or relaxing its charged groups (and also, to a lesser extent, the dipolar ones).[17,24,78,79,85] In the present work, the same trend for high ϵ_p values is consistently found for all dielectric profiles in Figures 1, 2, and 5 (except for single2 in Figure 1). That is, the inclusion of tautomers in truly titrable sites, alcohols or even water molecules does not affect the optimal value of $\epsilon_p \approx 20$. Thus, the suggestion made by Alexov and Gunner[23] that the inclusion of proton conformational isomerism makes possible to use lower ϵ_p values is not confirmed in the present work. This may be due to our simple set of tautomer geometries (see Computational Methods), or it may reflect the fact that more extensive conformational aspects must be included to lower the optimal ϵ_p value.

In any case, the particular value of 20 should not be seen as the "proper" value of ϵ_p in CE calculations, justified by some underlying theoretical reason; this value is just the one giving the best fit for the current CE methodology and set of conditions used here. Even restricting to the use of a single ϵ_p value (but see below), it is easy to devise modifications that may lead to different optimal ϵ_p values. Since this quantity largely reflects the large dielectric response due to the mobility of charged groups, the more mobile groups at the protein surface may be regarded as part of the solvent region, which already has a naturally high dielectric constant; this was done some time ago for the Tanford−Kirkwood model[87] and recently suggested for CE models.[78] A similar approach is to allow for a more extensive penetration of the water by assigning a smaller radius to the spherical probe used to compute the molecular surface (see Computational Methods). This results in a more extensive solvation of surface groups, which in principle should help to mimic their high dielectric response; consequently, a smaller ϵ_p value should be necessary. This is in fact observed, as shown in Figure 9. It is interesting that the form of the curves remains essentially unchanged; i.e., by changing the probe radius, we obtain a "rescaling" of ϵ_p. This clearly shows that the choice of

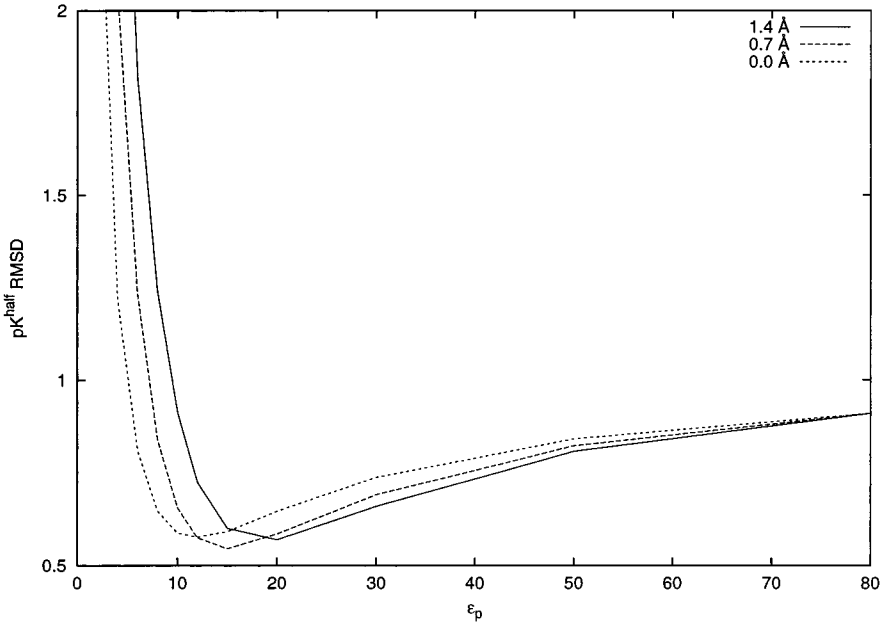

Figure 9. RMSD between experimental and calculated pK^{half} values using different radii for the water probe.

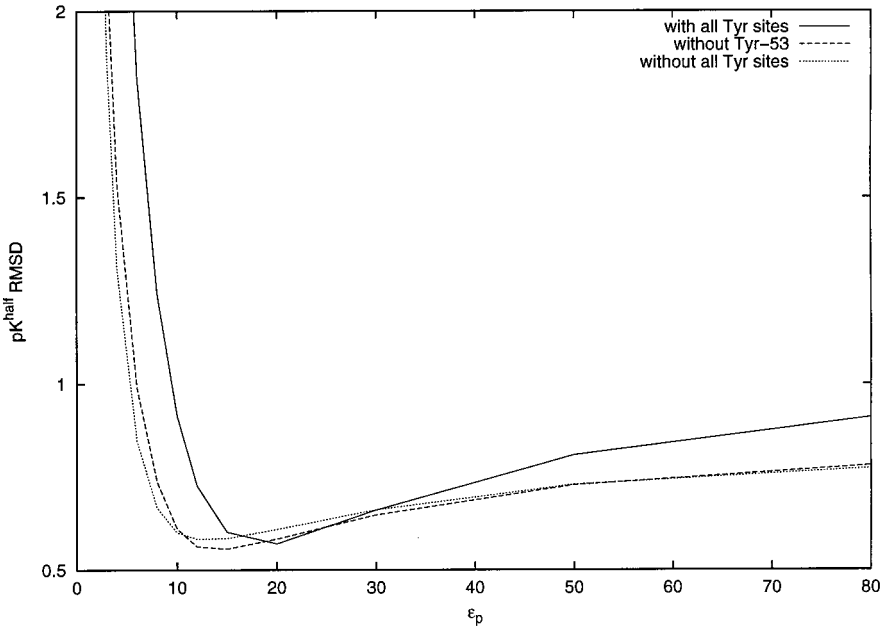

Figure 10. RMSD between experimental and calculated pK^{half} values with and without the inclusion of Tyr sites.

an optimal ϵ_p cannot be dissociated from the details of the CE model.

Finally, we note that, similar to what was observed by Demchuk and Wade,[21] our best pK^{half} predictions occur at different ϵ_p values for different sites (results not shown); a similar result was observed using the PDLD/S method.[79] This means that the choice of sites to be included in the calculation of the pK^{half} RMSD can affect the ϵ_p value that gives the best overall prediction. For example, given that the predicted pK^{half} of Tyr-53 is totally wrong at low ϵ_p values, we could think of excluding this site from the RMSD calculation; since this also holds, to a lesser extent, for other Tyr sites, we could in fact exclude all of them. Since these sites require high ϵ_p values to be reasonably predicted, the consequence of their removal is a global shift in the dielectric profile toward lower ϵ_p values, as shown in Figure 10. This results in a change of the RMSD value at $\epsilon_p = 4$ from 2.98 (with all Tyr) to 1.53 (without Tyr-53) and 1.31 (without all Tyr), which are much more similar to the values of the studies in Table 2 using $\epsilon_p = 4$, where such exclusions were done.[15,23] (The low RMSD from Beroza and Case[20] probably results from the inclusion of more extensive conformational effects.) These results show again the relative nature of the optimal ϵ_p in this case with respect to the choice of protonatable sites to be examined;[88] furthermore, due to a generally smaller reorganization, redox sites seem to require lower ϵ_p values than protonatable ones.[90,91] On the other hand, this issue naturally leads to the idea that the use of optimal local ϵ_p values could lead to better predictions because the local ϵ_p could in principle account for the particular conformational characteristics of each site (and also electronic polarizability,[92] though this effect would be much smaller). Unfortunately, there is no evident correlation between the optimal ϵ_p values and the static structural features of the site, even if some (weak) correlation with accessibility seems to exist.[21] Thus, it seems difficult to devise a criterion to decide whether a site is "low-dielectric" or "high-dielectric". In any case, even if such

classification could be done, the results of a calculation using different ϵ_p values for different regions could show different trends for the sites, since, for example, an optimal $\epsilon_p = 4$ obtained for a site using the same ϵ_p for the whole protein may loose its optimal nature when neighboring regions are assigned higher ϵ_p values. Hence, the eventual improvement that one may hope to result from the use of local ϵ_p values does not seem easy to obtain from a posteriori rules derived from the usual CE-based calculations using a single ϵ_p.

5. Conclusions

The present article shows that the energetics of proton isomerism in protein systems can be formally addressed using a thermodynamically equivalent system of nontautomeric sites. The treatment adopted here assumes that only the neutral form of the protonatable sites has tautomeric forms (as for usual sites), but other types of site can in principle be included if the reference state is changed accordingly.

The inclusion of tautomerism is shown to improve the pK^{half} computed values due to both energetic and entropic reasons. The method is also shown to be extremely robust to the inclusion of orientable water molecules, which further improves the predictions. In all cases, and in agreement with previous works,[17,21,26] we find a high vale of ϵ_p ($\sim$20) to give the best pK^{half} predictions. The number of tautomers per site was kept small in the present work, but it can easily be increased if needed.

The treatment of tautomerism proposed here can be easily extended to redox sites, namely, to the simulation of the joint equilibrium of protons and electrons.[31]

Acknowledgment. We thank Paulo J. Martel for his critical reading of the manuscript and Joaquim Mendes for helpful discussions. We are also grateful to the anonymous reviewers for their valuable comments. We acknowledge financial support from Fundação para a Ciência e a Tecnologia, Portugal, through Grants PRAXIS XXI/BPD/18899/98, PRAXIS XXI/P/BIO/14314/1998, and FCT 32789/99.

References and Notes

(1) Creighton, T. E. *Proteins*, 1st ed.; Freeman: New York, 1984.
(2) Fersht, A. R. *Enzyme Structure and Mechanism*, 2nd ed.; Freeman: New York, 1985.
(3) Perutz, M. F. *Science* **1978**, *201*, 1187.
(4) Linderstrøm-Lang, K. *C. R. Trav. Lab. Carlsberg* **1924**, *15*, 1.
(5) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.
(6) Tanford, C.; Roxby, R. *Biochemistry* **1972**, *11*, 2192.
(7) Warshel, A. *Biochemistry* **1981**, *20*, 3167.
(8) Warshel, A.; Russell, S. T. Q. *Rev. Biophys.* **1984**, *17*, 283.
(9) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219.
(10) Lee, F. S.; Chu, Z. T.; Warshel, A. *J. Comput. Chem.* **1993**, *14*, 161.
(11) Del Buono, G. S.; Figueirido, F. E.; Levy, R. M. *Proteins: Struct. Funct. Genet.* **1994**, *20*, 85.
(12) Bashford, D.; Gerwert, K. *J. Mol. Biol.* **1992**, *224*, 473.
(13) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045.
(14) Yang, A.-S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins: Struct. Funct. Genet.* **1993**, *15*, 252.
(15) Yang, A.-S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459.
(16) Oberoi, H.; Allewell, N. M. *Biophys. J.* **1993**, *65*, 48.
(17) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415.
(18) You, T. J.; Bashford, D. *Biophys. J.* **1995**, *69*, 1721.
(19) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35*, 7819.
(20) Beroza, P.; Case, D. A. *J. Phys. Chem.* **1996**, *100*, 20156.
(21) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373.
(22) Gibas, C. J.; Subramanian, S. *Biophys. J.* **1996**, *71*, 138.
(23) Alexov, E. G.; Gunner, M. R. *Biophys. J.* **1997**, *72*, 2075.
(24) Sham, Y. Y.; Chu, Z. T.; Warshel, A. *J. Phys. Chem. B* **1997**, *101*, 4458.
(25) Zhou, H.-X.; Vijayakumar, M. *J. Mol. Biol.* **1997**, *267*, 1002.
(26) Schaefer, M.; Sommer, M.; Karplus, M. *J. Phys. Chem. B* **1997**, *101*, 1663.
(27) Lancaster, C. R. D.; Michel, H.; Honig, B.; Gunner, M. R. *Biophys. J.* **1996**, *70*, 2469.
(28) Soares, C. M.; Martel, P. J.; Carrondo, M. A. *J. Biol. Inorg. Chem.* **1997**, *2*, 714.
(29) Kannt, A.; Lancaster, C. R. D.; Michel, H. *Biophys. J.* **1998**, *74*, 708.
(30) Martel, P. J.; Soares, C. M.; Baptista, A. M.; Fuxreiter, M.; Náray-Szabó, G.; Louro, R. O.; Carrondo, M. A. *J. Biol. Inorg. Chem.* **1999**, *4*, 73.
(31) Baptista, A. M.; Martel, P. J.; Soares, C. M. *Biophys. J.* **1999**, *76*, 2978.
(32) Sharp, K. A.; Honig, B. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 301.
(33) Rashin, A. A.; Bukatin, M. A. *Biophys. Chem.* **1994**, *51*, 167.
(34) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144.
(35) Nielsen, J. E.; Andersen, K. V.; Honig, B.; Hooft, R. W. W.; Klebe, G.; Vriend, G.; Wade, R. C. *Protein Eng.* **1999**, *12*, 657.
(36) Hereafter, we will designate the different protonation isomers of a site as tautomers. Strictly speaking, this term should not be used forstereoisomers,[37] as it is the case for carboxyl and amino sites, but it will be used here more freely as referring to any isomers which are in equilibrium with each other. Furthermore, we will extend the designation to include alcohol, thiol, and water "rotamers".
(37) Eliel, E. L.; Wilen, S. H. *Stereochemistry of Organic Compounds*; Wiley-Interscience: New York, 1994.
(38) Spassov, V.; Bashford, D. *J. Comput. Chem.* **1999**, *20*, 1091.
(39) Alexov, E. G.; Gunner, M. R. *Biochemistry* **1999**, *38*, 8253.
(40) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins: Struct. Funct. Genet.* **1997**, *27*, 523.
(41) Wlodek, S. T.; Antosiewicz, J.; McCammon, J. A. *Protein Sci.* **1997**, *6*, 373.
(42) Rabenstein, B.; Ullmann, G. M.; Knapp, E.-W. *Eur. Biophys. J.* **1998**, *27*, 626.
(43) van Vlijmen, H. W. T.; Schaefer, M.; Karplus, M. *Proteins: Struct. Funct. Genet.* **1998**, *33*, 145.
(44) Warshel, A.; Papazyan, A. *Curr. Opin. Struct. Biol.* **1998**, *8*, 211.
(45) Frölich, H. *Theory of Dielectrics*, 2nd ed.; Oxford University Press: Oxford, 1958.
(46) Kuramitsu, S.; Hamaguchi, K. *J. Biochem.* **1980**, *87*, 1215.
(47) Bartik, K.; Redfield, C.; Dobson, C. M. *Biophys. J.* **1994**, *66*, 1180.
(48) Beroza, P.; Fredkin, D. R.; Okamura, M. Y.; Feher, G. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 5804.
(49) Gilson, M. K. *Proteins: Struct. Funct. Genet.* **1993**, *15*, 266.
(50) Hill, T. L. *Statistical Mechanics*; McGraw-Hill: New York, 1956.
(51) Hodsdon, J. M.; Brown, G. M.; Sieker, L. C.; Jensen, L. H. *Acta Crystallogr., B* **1990**, *46*, 54.
(52) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B., Jr.; E. F. M.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Biol.* **1977**, *112*, 535.
(53) Eisenhaber, F.; Argos, P. *J. Comput. Chem.* **1993**, *14*, 1272.
(54) Eisenhaber, F.; Lijnzaad, P.; Argos, P.; Sander, C.; Scharf, M. *J. Comput. Chem.* **1995**, *16*, 273.
(55) Guex, N.; Peitsch, M. C. *Electrophoresis* **1997**, *18*, 2714.
(56) van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS) Library Manual*; Biomos: Groningen, The Netherlands, 1987.
(57) Smith, L. J.; Mark, A. E.; Dobson, C. M.; van Gunsteren, W. F. *Biochemistry* **1995**, *34*, 10918.
(58) Janin, J.; Wodak, S.; Levitt, M.; Maigret, B. *J. Mol. Biol.* **1978**, *125*, 357.
(59) Gelin, B. R.; Karplus, M. *Biochemistry* **1979**, *18*, 1256.
(60) MacArthur, M. W.; Thornton, J. M. *Acta Crystallogr., D* **1999**, *55*, 994.
(61) Bashford, D. An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules. In *Scientific Computing in Object-Oriented Parallel Environments*; Ishikawa, Y., Oldehoeft, R. R., Reynders, J. V. W., Tholburn, M., Eds.; ISCOPE97; Springer: Berlin, 1997; p 233.
(62) Richards, F. M. *Annu. Rev. Biophys. Bioeng.* **1977**, *6*, 151.
(63) Gilson, M. K.; Sharp, K. A.; Honig, B. *J. Comput. Chem.* **1987**, *9*, 327.
(64) Tanokura, M. *Biochim. Biophys. Acta* **1993**, *742*, 576.
(65) Antosiewicz, J.; Porschke, D. *Biochemistry* **1989**, *28*, 10072.
(66) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.
(67) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Clarendon: Oxford, 1987.
(68) Hastings, W. K. *Biometrika* **1970**, *57*, 97.
(69) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978.

(70) Jayaram, B.; Fine, R.; Sharp, K.; Honig, B. *J. Phys. Chem.* **1989**, *93*, 4320.
(71) Levy, R. M.; Belhadj, M.; Kitchen, D. B. *J. Chem. Phys.* **1991**, *95*, 3627.
(72) Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206.
(73) Papazyan, A.; Warshel, A. *J. Chem. Phys.* **1997**, *107*, 7975.
(74) Levy, R. M.; Gallicchio, E. *Annu. Rev. Phys. Chem.* **1998**, *49*, 531.
(75) Gilson, M. K.; Honig, B. H. *Biopolymers* **1986**, *25*, 2097.
(76) Harvey, S. C. *Proteins: Struct. Funct. Genet.* **1989**, *5*, 78.
(77) Warshel, A.; Åqvist, J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267.
(78) Simonson, T.; Perahia, D. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 1082.
(79) Sham, Y. Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744.
(80) Hill, T. L. *J. Phys. Chem.* **1956**, *60*, 253.
(81) Rees, D. C. *J. Mol. Biol.* **1980**, *141*, 323.
(82) Matthew, J. B. *Annu. Rev. Biophys. Biophys. Chem.* **1985**, *14*, 387.
(83) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671.
(84) King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366.
(85) Smith, P. E.; Brunne, R. M.; Mark, A. E.; van Gunsteren, W. F. *J. Phys. Chem.* **1993**, *97*, 2009.
(86) Simonson, T.; Perahia, D.; Bricogne, G. *J. Mol. Biol.* **1991**, *218*, 859.
(87) States, D. J.; Karplus, M. *J. Mol. Biol.* **1987**, *197*, 122.
(88) As noted by one reviewer, the ionization of Tyr-53 is associated with the alkaline denaturation of lysozyme.[89] Thus, by including this site in the RMSD calculation, we are assuming that the ϵ_p parameter should mimic the local denaturation needed to accommodate the ionized form. Alternatively, we may decide that ϵ_p should not be pushed that far and eliminate Tyr-53 from the RMSD calculation; in that case, the second curve of Figure 10 should be used to estimate the optimal ϵ_p. In any case, the optimal value remains in the 15−20 region.
(89) Imoto, T.; Johnson, L. N.; North, A. T. C.; Phillips, D. C.; Rupley, J. A. In *The Enzymes*, 3rd ed.; P. D. Boyer, Ed.; Academic Press: New York, 1972; Vol. 7, p 665−868.
(90) Muegge, I.; Qi, P. X.; Wand, A. J.; Chu, Z. T.; Warshel, A. *J. Chem. Phys. B* **1997**, *101*, 825.
(91) Sharp, K. A. *Biophys. J.* **1998**, *73*, 1241.
(92) Sharp, K. A.; Jean-Charles, A.; Honig, B. *J. Phys. Chem.* **1992**, *96*, 3822.