PROTEINS: Structure, Function, and Genetics 15:266–282 (1993)

Multiple-Site Titration and Molecular Modeling: Two Rapid Methods for Computing Energies and Forces for Ionizable Groups in Proteins

Michael K. Gilson
Department of Chemistry, University of Houston, Houston, Texas 77204-5641

ABSTRACT **Computer models of proteins frequently treat the energies and forces associated with ionizable groups as if they were purely electrostatic. This paper examines the validity of the purely electrostatic approach, and concludes that significant errors in energies can result from the neglect of ionization changes. However, a complete treatment of ionizable groups presents substantial computational obstacles, because of the large number of ionization states which must be examined in systems having multiple interacting titratable groups. In order to address this problem, two novel methods for treating the energetics and forces associated with ionizable groups with a minimum of computer time have been developed. The most rapid method yields approximate energies by computing the free energy of a single highly occupied ionization state. The second method separates ionizable groups into clusters, and treats intracluster interactions exactly, but intercluster interactions approximately. This method yields both accurate energies and fractional charges. Good results are obtained in tests of both methods on proteins having has many as 123 ionizable groups. The more rapid method requires computer times of 0.01 to 0.34 sec, while the more accurate method requires 0.7 to 15 sec. These methods may be fast enough to permit the incorporation of ionization effects in iterative computations, such as energy minimizations and conformational searches.**

Key words: electrostatics, conformational energies, solvation, ionization, force fields, energy functions

INTRODUCTION

Various methods exist for computing the electrostatic energy of a polypeptide or protein of specified three-dimensional conformation. These include the use of empirical dielectric screening functions, such as the distance-dependent dielectric constant $\epsilon = R$ and the more complex function of Mehler and Eichele[1]; the Tanford–Kirkwood[2] and modified Tanford–Kirkwood[3] models; the "protein-dipoles Langevin-dipoles" (PDLD) approach[4]; and detailed solutions of the Poisson equation[5,6] and of the Poisson–Boltzmann equation.[7–10] The use of improved electrostatics models should substantially improve the predictive power of molecular simulations and structural predictions.

However, the fact that the charges of ionizable groups are not fixed adds an additional level of complexity to the problem of protein energetics. Consider, for example, the case of a lysine side chain which is transferred from solution into a protein interior where it is highly desolvated. If the desolvation penalty is large enough, the lysine will deprotonate and become electrically neutral. However, few if any molecular force fields account for this phenomenon. For example, the hydration shell model[11,12] provides for approximately a 60 kcal/mol energy penalty for desolvation of an ammonium group, and a more recent energy function[13] allows about 70 kcal/mol for the desolvation of a primary amine. As discussed in detail below, such values overestimate the correct desolvation penalty by about a factor of 6, because they are based upon the assumption that the group remains ionized. Similarly, molecular mechanics simulations and conformational energy calculations usually presume each ionizable group to exist in a single fixed charge state, whose selection is based upon the pK_a of the group in isolation in solvent, rather than in its actual location in the macromolecule. As described below, this approach tends to yield energies which are too positive, when compared with energies based upon the full theory of multiple titrating groups.

Probably the chief reason that molecular models often set aside the full treatment of ionization equilibria is the fact that the computational obstacles can be quite challenging. This is a consequence of the fact that formal evaluation of pH-dependent charges and energies for a system of n ionizable groups requires enumeration of 2^n ionization states.

Received May 4, 1992; revision accepted July 22, 1992.
Address reprint requests to Dr. Michael K. Gilson, Department of Chemistry, University of Houston, Houston, TX 77204-5641.

This number is prohibitively large for many systems of interest. Various approximations have therefore been used to reduce the complexity of the problem. One well known approximation is the iterative method of Tanford and Roxby.[14] Here each ionizable group i feels each other group j according to its average charge, $<q_j>$. This method has recently been demonstrated to constitute a mean field approximation, which provides accurate results only when the ionization states of groups are negligibly correlated with each other.[15] The reduced site method[15] is a more precise alternative which involves identifying those groups which are essentially fully ionized or neutral, no matter what the charge states of other groups may be, and then eliminating these groups from the enumeration of states. Although this significantly speeds calculations, it still does not permit many systems of interest to be treated with reasonable speed. A powerful Monte Carlo method has also been developed,[16] which makes it possible to address very large systems quite efficiently. However, this method is still too slow to be useful in iterative applications, and there would clearly be a role for more rapid methods.

The present paper discusses the basic theory of multiple site titration, and provides expressions for the pH-dependent forces associated with ionizable groups. The significance of ionization energetics for molecular models is examined in calculations on model systems. Two novel, computationally rapid, methods for calculating the charges and energies of ionizable groups are presented. Tests of these methods for systems of up to 123 groups demonstrate their validity. It should be noted that the present work does *not* address the problem of calculating electrostatic interactions. Rather, the work is limited to the incorporation of such interactions into charge, energy and force calculations for ionizable groups.

METHODS

Ionization Polynomial

The theory presented here for the energetics and forces of ionizable groups is a straightforward application of the binding polynomial treatment of multiple ligand binding.[17] The free energy obtained is that for the hypothetical process going from the neutral form of a protein to the ionized protein at equilibrium with a solution of some pH. This free energy will be termed the "ionization energy." It is worth mentioning that the ionization energy is always less than or equal to zero, because it is defined as the free energy of a spontaneous ionization process which proceeds to equilibrium. The ionization energy is not purely electrostatic, because it also depends implicitly upon the chemical potential of the proton, and the bonding energy of the proton to the ionizable groups. However, for convenience the ionization energy may temporarily be termed a component of the electrostatic energy of the protein.

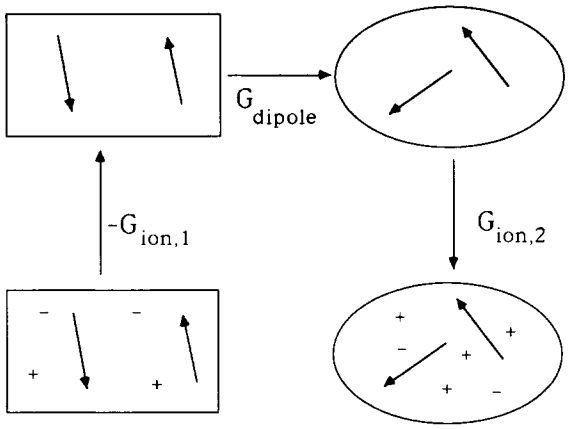

Fig. 1. Thermodynamic pathway for definition of the ionization energy. On the left, a protein in conformation 1, in equilibrium with solvent at some pH, is neutralized ($-G_{ion,1}$). In the next step, it is rearranged to conformation 2 (G_{dipole}), and finally it is permitted to ionize once again ($G_{ion,2}$). Plus and minus signs indicate ionized groups. Arrows indicate a permanent charge distribution.

With this provisional definition, the change in electrostatic free energy upon going from protein conformation 1 to conformation 2 can be analyzed in terms of the thermodynamic pathway shown in Figure 1. Starting from a state where it is in equilibrium with a solution having some pH, conformation 1 is neutralized, with an ionization energy change of $-G_{ion,1}$. The conformation of the neutral protein is then changed to conformation 2, with a change in electrostatic energy which may be termed the "dipole" term, G_{dipole}, because there are no groups bearing a net charge. This term is obtained by the straightforward application of an electrostatics model to the neutral form of each conformation. Finally, conformation 2 is allowed to ionize, yielding a free energy of $G_{ion,2}$. The difference $G_{ion,2} - G_{ion,1}$ results entirely from the changes in the electrostatic environments of the ionizable groups as a consequence of the conformational change. The net "electrostatic" energy change of the completed conformational change is $G_{ion,2} - G_{ion,1} + G_{dipole}$.

We wish to derive an expression for the free energy difference between a polypeptide chain in a hypothetical neutral form and the same chain in equilibrium with a solution of arbitrary pH. Because the following derivation closely follows that of Schellman,[17] it is presented here in a somewhat abbreviated form. We consider a dilute solution of a protein at concentration P having n titratable groups (1, 2, 3, . . . , i, . . . , n). The groups are described by a vector $\mathbf{z}(i)$ which is $+1$ for each base and -1 for each acid. At any moment, each ionizable group is considered to be either fully neutral or fully charged. Thus the protein fluctuates among a set of 2^n possible ionization forms α. Each ionization form

is described by a vector $\mathbf{x}_\alpha$ such that $x_\alpha(i) = 1$ if group i is ionized in state α, and $x_\alpha(i) = 0$ if group i is neutral in state α.

The concentration of the neutral form of the protein ($\alpha = 0$) is P_0. The concentration of ionization form α of the protein is P_α, and the fractional occupancy of ionization form α is $f_\alpha = P_\alpha/P$. The fractional occupancy of the neutral form is $f_0 = P_0/P$. The net charge of ionization form α is

$$q_\alpha = \sum_{i=1}^{n} x_\alpha(i) z(i). \tag{1}$$

The mean charge of each protein molecule at equilibrium is equal to the mean number of protons which bind to the neutral protein on going to equilibrium, and may be written

$$q = \sum_\alpha f_\alpha q_\alpha. \tag{2}$$

Finally, we define the equilibrium constants K_α for the formation of ionization form α from the neutral protein. These may be written

$$K_\alpha = \left(\frac{P_\alpha}{P_0[H^+]^{q_\alpha}}\right)_{\text{equilibrium}} = e^{-\beta(\mu_\alpha^0 - \mu_0^0 - q_\alpha \mu_{H^+}^0)} \tag{3}$$

where $\beta = 1/RT$, and μ_a^0, μ_0^0, and $\mu_{H^+}^0$ are the standard chemical potentials of, respectively, the protein in ionization form α, the neutral protein, and the proton.

In the protein's neutral state, where $P_0 = P$, the free energy of the system will be

$$G_I = \mu_0^0 + RT \ln P_0 + q(\mu_{H+}^0 + RT \ln [H^+]) + G_{ex}. \tag{4}$$

G_{ex} represents the free energy of components of the system not expressly included in the formula. The third term here represents the free energy of the average net number of protons which will bind to the protein when it is allowed to equilibrate with the solvent.

At equilibrium with a solution of a given pH, the protein will take on some distribution of the ionization states α. The free energy of the protein under these circumstances is then given by

$$G_{II} = f_0(\mu_0^0 + RT\ln P_0) + \sum_{\alpha \neq 0} f_\alpha(\mu_\alpha^0 + RT\ln P_\alpha) + G_{ex}. \tag{5}$$

Using Eq. (2) and the relation

$$f_0 = 1 - \sum_{\alpha \neq 0} f_\alpha \tag{6}$$

it can be shown that

$$G_{II} - G_I = RT\ln(P_0/P) + \sum_{\alpha \neq 0} f_\alpha [\mu_\alpha^0 + RT\ln(P_\alpha/P_0) - \mu_0^0 - q_\alpha(\mu_{H^+}^0 + RT\ln[H^+])]. \tag{7}$$

Using Eq. (3), the contents of the square brackets goes to zero for each α, leaving the ionization energy

$$G_{ion} = G_{II} - G_I = RT \ln (P_0/P) \equiv -RT \ln \Sigma \tag{8}$$

where Σ, a slightly modified binding polynomial which may be termed the ionization polynomial, is given by

$$\Sigma \equiv P/P_0 = 1 + \sum_{\alpha \neq 0} K_\alpha [H^+]^{q_\alpha}. \tag{9}$$

We expect that the equilibrium constants K_α will be different for two different conformations of a polypeptide. The above derivation makes it possible to go from the values of K_α for the two conformations, to the difference in the ionization free energy

$$G_{ion,2} - G_{ion,1} = -RT\ln \frac{\Sigma_2}{\Sigma_1} \tag{10}$$

where $G_{ion,1}$ and $G_{ion,2}$ are the values of the ionization energy, $G_{II} - G_I$, for conformations 1 and 2, and Σ_1 and Σ_2 are the values of the ionization polynomial for conformations 1 and 2.

It is now necessary to connect the ionization equilibria of the individual ionizable groups with the equilibrium constants K_α. Since G_{ion} represents the free energy of ionizing the protein, the corresponding equilibria will be, for acidic groups,

$$A_iH \rightleftharpoons A_i^- + H^+ \tag{11}$$

with equilibrium constant K_{ai}; and for basic groups

$$B_i + H^+ \rightleftharpoons B_iH^+ \tag{12}$$

with equilibrium constant K_{ai}^{-1}. Defining the K_{ai} values to correspond to pK_as of ionizable groups unperturbed by the protein environment we have

$$K_\alpha = \prod_{i=1}^{n} K_{ai}^{-x_\alpha(i)z(i)}. \tag{13}$$

More generally, the pK_a of a group will be perturbed by desolvation effects, interactions with dipolar groups, and interactions with other ionizable groups. The extra free energy, relative to the unperturbed state, of ionizing group i when all other ionizable groups are taken to be neutral will be termed G_i. (Note that G_i is responsible for the difference between the group's intrinsic pK_a, as customarily defined, and its pK_a in the absence of any perturbing influence.) The extra free energy of ionizing group i when a group j is ionized will be G_{ij}. Taking higher order interactions (e.g., G_{ijk}) to be negligible, we can write

$$K_\alpha = e^{-\beta \sum_{i=1}^{n} x_\alpha(i) \left[G_i + \sum_{j>i}^{n} x_\alpha(j) G_{ij} \right]} \prod_{i=1}^{n} K_{ai}^{-x_\alpha(i)z(i)}. \tag{14}$$

The summation in the exponent of e is $G_{elec,\alpha}$, the

purely electrostatic energy of ionization state α. Defining, for notational convenience.

$$A_i = K_{ai}^{-z(i)}[H^+]^{z(i)} \tag{15}$$

we see that

$$\begin{aligned} G_{\text{ion}} &= -RT\ln\Bigg(1 + \sum_i^n A_i e^{-\beta G_i} \\ &\quad + \sum_i^n \sum_{j>i}^n A_i A_j e^{-\beta(G_i+G_j+G_{ij})} + \cdots \\ &\quad + e^{-\beta\sum_i^n \left(G_i + \sum_{j>i}^n G_{ij}\right)} \prod_i^n A_i\Bigg) \\ &= -RT\ln\left(\sum_{\alpha=0}^{2^n-1} e^{-\beta G_{\text{elec},\alpha}} \prod_i^n A_i^{\chi_\alpha(i)}\right). \end{aligned} \tag{16}$$

It is perhaps worth making explicit that the energies G_i and G_{ij} are here assumed to be independent of ionization state, and to be purely electrostatic in origin, although the latter assumption is not critical.

Forces on Ionizable Groups

Certain calculations require that expressions for interatomic forces be available. For example, energy minimizations cannot be performed if the energy function in use is not expressed in terms of a "force field" relating atomic coordinates to interatomic forces. In order to facilitate the incorporation of ionization energetics into such calculations, the present section derives expressions for the forces which correspond to the ionization free energy.

In the interests of notational simplicity, consider the case of a molecule having only two ionizable groups, i and j, with associated electrostatic energies G_i, G_j, and G_{ij}, as defined above. The ionization energy associated with an arbitrary conformation will be given by

$$\begin{aligned} G_{\text{ion}} &= -RT \ln \Sigma \\ &= -RT \ln[1 + A_i e^{-\beta G_i} + A_j e^{-\beta G_j} \\ &\quad + A_i A_j e^{-\beta(G_{ij} + G_i + G_j)}]. \end{aligned} \tag{17}$$

Any atom whose movement influences this energy will feel a corresponding force. If y represents some internal coordinate of the molecule, then the force acting along this coordinate proves to be

$$\begin{aligned} F_y &= -\frac{\partial G_{\text{ion}}}{\partial y} \\ &= -\langle x(i)\rangle \frac{\partial G_i}{\partial y} - \langle x(j)\rangle \frac{\partial G_j}{\partial y} - \langle x(i)x(j)\rangle \frac{\partial G_{ij}}{\partial y}. \end{aligned} \tag{18}$$

This force is simple the average over ionization states of the purely electrostatic forces acting on the atom. This result applies as well to larger systems.

As an example, if G_{ij} is assumed to be based upon Coulomb's law, so that

$$G_{ij} = \frac{z(i)z(j)}{4\pi\epsilon r_{ij}} \tag{19}$$

where ε is a permittivity, and if r_{ij} does not influence G_i or G_j, then the force acting between atoms i and j may be determined by setting y in the above expressions to r_{ij}, yielding

$$F_{ij} = -\langle x(i)x(j)\rangle \frac{z(i)z(j)}{4\pi\epsilon r_{ij}^2}. \tag{20}$$

These expressions demonstrate that although the ionization *energy* of a system is not equal to the electrostatic energy averaged over ionizations states, as previously noted,[18] the *forces* associated with ionizable groups are, in effect, purely electrostatic.

Predominant State Approximation

Although the full evaluation of an ionization energy requires the enumeration of each possible ionization state, it is demonstrated here that the free energy of a single highly populated ionization state will frequently represent an excellent estimate of the overall ionization energy. A rapid method for guessing a highly populated state is then described. Tests of the method are described in a subsequent section.

The ionization polynomial may be written compactly as

$$G_{\text{ion}} = -RT\ln\sum_{\alpha=0}^{2^n-1} e^{-\beta G_\alpha} \tag{21}$$

where G_α is the free energy of ionization state α relative to the neutral state. Defining α_p to be our guess for the predominant state, and r_α to be the ratio of the population of state α to that of state α_p,

$$\begin{aligned} G_{\text{ion}} &= -RT\ln\left(e^{-\beta G_{\alpha_p}} \sum_{\alpha=0}^{2^n-1} r_\alpha\right) \\ &= G_{\alpha_p} - RT\ln\sum_{\alpha=0}^{2^n-1} r_\alpha. \end{aligned} \tag{22}$$

It is easy to show that the fractional occupany of state α_p, f_p, is equal to $(\sum_{\alpha=0}^{2^n-1} r_\alpha)^{-1}$. Thus

$$G_{\text{ion}} = G_{\alpha_p} + RT \ln f_p. \tag{23}$$

This expression is exact. In the simplest predominant state method, G_{ion} can be approximated as equal to G_{α_p}. What is perhaps surprising is the fact that even if state α_p is not very highly occupied, the error due to this approximation ($RT \ln f_p$) can be quite small. For example, if the true occupancy of the guessed state is 0.25, RT ln (0.25) is only −0.83 kcal/mol. If the occupancy is only 0.1, the error is still only −1.4 kcal/mol. Thus even a rather poor guess for the pre-

dominant state can yield a good guess for the ionization energy. Moreover, if an estimate of f_p can be generated, the error can be reduced.

The particular method proposed here for guessing the predominant state is a variant of the iterative Tanford and Roxby[14] method for computing titration curves. The chief modification is that each group is forced to be either neutral or ionized at all times—no fractional ionizations are allowed. Each group begins with an intrinsic pK_a, based upon the pK_a of a model compound, modified by the self energy (G_i) of the group. The fractional ionization of each group is then set to either 0 or 1, depending solely upon whether its pK_a is above or below the current pH, and whether it is an acid or base. The pK_a of group 1 in the electrostatic field of all the other groups is then recomputed, and its charge reassigned according to the same criterion. This procedure is repeated for each group in the list, and is iterated until the charges stop changing, and a single ionization state has been selected. The free energy of this state is evaluated. Then the iterative procedure is repeated from the beginning, but running through the list of groups in the opposite direction (n to 1). This provides a second guess for the predominant state. Of the two states thus generated, the one with the lower free energy is selected as the predominant one, and its free energy is taken to be the best guess.

The two runs also yield two different sets of pK_as. These are used to estimate f_p, as follows. For each group i, the pK_{a_i} from each run is used to compute a fractional ionization θ_i, using the standard expression for independent groups.[14] The fractional ionizations from the two runs are averaged, to provide a set of best-guess fractional ionizations. The deviation of each fractional ionization from the ionization state to which it is closest (0 or 1) may be written as

$$d_i = 0.5 + |\theta_i - 0.5|. \tag{24}$$

Assuming statistical independence, the fractional occupancy of the predominant state is then estimated as

$$f_p \approx \prod_i d_i. \tag{25}$$

The importance of this correction is evident in the case of a system containing say n noninteracting groups of equal pK_a, at pH = pK_a. Under these circumstances, each of the 2^n possible ionization states, including the neutral state, is equally occupied. Therefore, the uncorrected predominant state approximation will yield an incorrect ionization energy of 0, because each ionization state has free energy equal to that of the neutral state. However, the correction term in this case is exact, because $\Pi_i d_i$ is exactly equal to f_p, and therefore the correct ionization energy, $-nRT$ ln 2, will be obtained from the three previous equations.

Although, as shown below, this method yields remarkably accurate energies, the charge estimates are less satisfactory, and are not further analyzed in this work. However, the following section describes a novel method for rapidly calculating both energies and fractional charges with considerable accuracy.

Cluster Method

This method reduces the number of ionization states which need to be enumerated by separating the ionizable groups into clusters which, although they may have significant intracluster charge–charge correlations, have small intercluster correlations. A full ionization polynomial is evaluated for each cluster, but cluster–cluster interactions are treated by the mean field approximation.[15] A more limited hybridization of the mean field method with exact calculations has been suggested previously,[15] and a recent review mentions the use of a similar approach in unpublished work.[19] The method will be successful to the extent that ionizable groups in real proteins actually can be subdivided into clusters of manageable size, having small cluster–cluster charge correlations. The theory is presented next, followed by the description of criteria and a method for separating groups into clusters.

In the mean field approximation, the ionization equilibrium of group i is influenced by group j according to $\theta_j G_{ij}$. In effect, the pK_a of group i is perturbed according to the mean electrostatic potential at i due to j, without regard for possible charge–charge correlations between the two groups. Thus, if we focus on a particular cluster, I, having n_I ionizable groups, the ionization polynomial for that cluster can be written as

$$\Sigma_I = \sum_{\alpha=0}^{2^{n_I}-1} \left(e^{-\beta G_{\text{elec},\alpha}} e^{-\beta \sum_{i=1}^{n_I} x_\alpha(i) \sum_{k \ni I} \theta_k G_{ik}} \prod_{i=1}^{n_I} A_i^{x_\alpha(i)} \right) \tag{26}$$

where the outer sum is over the ionization states of cluster I only, and the expression $k \ni I$ signifies that k ranges over the indices of all clusters other than I. A similar expression will be used for each cluster.

The method is implemented by initially setting each group to its fully ionized state. (This choice of initial state is dictated by the fact that the mean field method can miss the cooperative ionization of two otherwise neutral groups if the groups are initialized to the neutral state.) The ionization polynomial of cluster I is then evaluated, using the initial-guess charges for groups in other clusters, and is used to update the fractional charges of the groups in cluster I, using the formula

$$\theta_i = \frac{1}{\Sigma_I} \sum_{\alpha=0}^{2^{n_I}-1} x_\alpha(i) e^{-\beta G_{\text{elec},\alpha}} e^{-\beta \sum_{i=1}^{n_I} x_\alpha(i) \sum_{k \ni I} \theta_k G_{ik}} \prod_{i=1}^{n_I} A_i^{x_\alpha(i)} \tag{27}$$

which simply states that θ_i equals the summed fractional occupancies of all the ionization states having group i ionized, i.e., with $x(i) = 1$. These new charges are used in the computation of the fractional charges of the next cluster, using its own independent set of ionization states. All clusters are thus updated, and the process is iterated to self-consistency, where the criterion for convergence is that all fractional charge changes fall below some small value. This procedure is the same as that of Tanford and Roxby[14] except that the iterative loop is over clusters of groups, rather than individual groups. In fact, the Tanford and Roxby method is recovered when each cluster consists of only one group. In order to force convergence of this potentially oscillatory iterative system, the fractional charges of each cluster actually feel the fractional charges of the other clusters averaged over the previous two iterations.

Although the iterations proceed reasonably quickly for clusters of up to about 10 groups, they quickly become slow when larger clusters are used. Convergence can be accelerated by "preprocessing" with clusters of maximum size say 10, then regenerating the larger clusters, and iterating to convergence starting with the fractional charges already generated.

It might be imagined that the ionization energy of a system of clusters treated in this way would be simply the sum of the ionization energies of the clusters, calculated using their ionization polynomials defined in Eq. (26). This proves to be incorrect, however, because it double counts the electrostatic cluster–cluster interactions. This double-counting, and the requisite correction, may be understood by the following argument, illustrated in Figure 2.

We consider two clusters, I and II, which will go from their hypothetical neutral states to their ionized states in each other's presence. We wish to compute the free energy of this ionization process. We begin by allowing each cluster to ionize in the presence of a set of *fixed* charges q_k, having values and locations equal to those which the other cluster will have when fully ionized at the end of the process [$q_k = \theta_k z(k)$]. The free energy of ionizing the clusters under these conditions is given by

$$G_1 = -RT \ln \Sigma_{\mathrm{I}} - RT \ln \Sigma_{\mathrm{II}}. \qquad (28)$$

The two clusters are now ionized, each in the presence of an artificial set of fixed charges. The process is completed by removing the sets of fixed charges to infinite distance, and putting cluster I in its correct location relative to cluster II. Because clusters I and II now feel the same electrostatic potentials as they did in the presence of the fixed charges (based upon the mean field assumption), their distributions of ionization states, and therefore energies, are unchanged. However, in separating each cluster from a set of fixed charges, there is an electrostatic energy

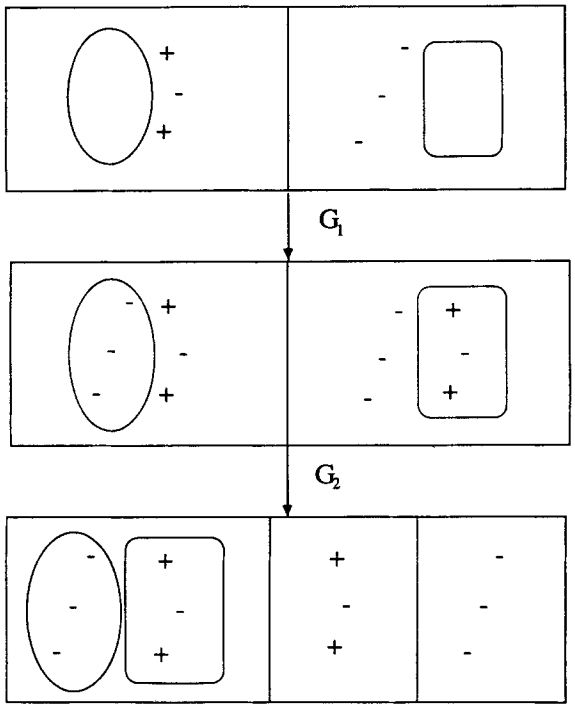

Fig. 2. Thermodynamic pathway for analyzing energetics of cluster method. **Top panel:** clusters I (oval) and II (rectangle) in neutral form, each in the presence of a set of fixed charges representing the fractional charges that the other cluster will have at equilibrium. **Middle panel:** clusters I and II ionized in the presence of the fixed charges, in equilibrium with a solvent having some pH. **Lower panel:** clusters I and II interacting with each other, fixed charges removed to infinity.

change of $-2\Sigma_i\Sigma_k\theta_i\theta_k G_{ik}$, where i indexes cluster I and k indexes cluster II. Moreover, in reassembling the two clusters, there is another electrostatic energy change of $\Sigma_i\Sigma_k\theta_i\theta_k G_{ik}$. Thus, the second step of the process involves a net energy change of

$$G_2 = -\sum_i \sum_k \theta_i\theta_k G_{ik}. \qquad (29)$$

The final expression for ionization energy is therefore

$$G_{ion} = G_1 + G_2 = -RT \ln \Sigma_I - RT \ln \Sigma_{II} - \sum_i \sum_k \theta_i\theta_k G_{ik}. \qquad (30)$$

This analysis can be generalized for larger systems, the final equation taking on the same form. Appendix A provides an alternative derivation based upon the variational formulation of Bashford and Karplus.[15]

It is evident that the criteria used to separate groups into clusters are of central importance in this approach. In the present work, two criteria are used. First, as in the reduced site method,[15] the maximal and minimal possible fractional charge of each group i is calculated for the pH of interest, by setting the charges of all other groups to their minimal and max-

imal values, respectively. If the maximal and minimal fractional charges of the group are both within some cutoff value (usually 0.05 protons) of either 0 or 1.0, the group is designated a "fixed" group, and is placed in a cluster by itself. Such groups may have strong interactions with other groups, but because there is no way their fraction charges can vary much, they can be treated accurately by the mean-field method. Note that this treatment of "fixed" groups here differs from that of the reduced site method. This is because the groups are not truly fixed, but are treated according to the mean field method as single-group clusters. This refinement follows a suggestion made previously by Bashford and Karplus.[15]

The separation of the remaining groups into clusters with small charge–charge correlations is based upon the presumption that charge–charge correlations will be small if charge–charge interaction energies are small. A recursive search method is used to establish clusters of groups whose charge–charge interactions with groups in other clusters are all below some cutoff value. Initially, this cutoff energy (G_c) is set to zero. A search is begun with the first non-fixed group (group 1) in the molecule. Any group i which has an interaction energy G_{i1} with group 1 of greater than G_c is added to the growing cluster. Any group j which in turn has an interaction energy $G_{ij} > G_c$ with group i is added to the same cluster. This procedure is continued until every group linked to group 1 by groups having mutual interaction energies greater than G_c has been added to the cluster. This establishes the first cluster. If this cluster proves to be larger than some established size limit (say 10 groups), then the value of G_c is incremented by 0.1 kcal/mol, and group 1 is again used to initiate a cluster. G_c is incremented until the cluster is within the allowed size limit. A similar search is now started using the next group which does not already belong to a cluster. In the end, every group in the molecule will have been assigned to a cluster of reasonable size. It is evident that the smaller the values of G_c needed to establish clusters, the more accurate the cluster method will be, because all intercluster interactions are treated by the mean field theory, and the G_c values represent the maximum intercluster interaction energies. Note that the required values of G_c will be pH- and system-dependent.

Numerical Tests

The approximate methods described here are tested on a series of randomly generated systems of ionizable groups, and on five proteins of known three-dimensional structure. The reported ionization energies in these calculations are all referred to the ionization energies of the same groups in the absence of electrostatic interactions, and are therefore really ΔG_{ion} values. These values can be viewed as approximations to the ionization energy change upon folding, to the extent that electrostatic interactions are small in the unfolded state. All the computations are performed on either a Silicon Graphics Iris 4D/320VGX or a Silicon Graphics Iris 4D/GTX. In timing tests with the code used here, the latter runs at very close to half the speed of the former. For uniformity, therefore, all reported CPU timings are corrected to values for the 320VGX system. It should be noted that little effort has been made to maximize the speed of the computer programs for the predominant state and cluster methods.

For the lysozyme calculations, the intrinsic pK_as and charge–charge interaction energies are those computed by Bashford and Karplus for the tetragonal crystal form of hen egg white lysozyme using the finite-difference Poisson–Boltzmann method.[8] However, as discussed below, all the interactions energies are multiplied by a factor of 3 to create a more challenging test case. Because arginines are treated as fixed charges, this system contains 21 ionizable groups.

For the other protein calculations, the group pK_as are as follows: Asp 4.0, Glu 4.4, Lys 10.4, Arg 12.5, Tyr 10.2, His 6.3, N terminus 7.5, C terminus 3.8, heme propionate 4.0. Carboxylate charges are placed on the carboxylate carbon, ammonium charges on the nitrogen. Arginine's charge is placed on CZ, and tyrosine's on OH. The only thiol included is that of Cys-247 of rhodanese, which is assigned a pK_a of 8.3. The charge is placed on the sulfur. Histidine's charge is arbitrarily assigned to atom NE2. The protein data bank[20] is the source of the atomic coordinates of the proteins. The proteins examined are poplar plastocyanin[21] (6PCY; 99 residues, 27 ionizable groups), sperm whale deoxymyoglobin[22] (1MBD; 153 residues, 62 ionizable groups), rhodanese[23] (1RHD; 293 residues, 93 ionizables), and pig heart citrate synthase[24] (4CTS, A chain; 437 residues, 123 ionizables). The copper ions of plastocyanin, the iron of myoglobin's heme, and the oxaloacetate complexed with citrate synthase are ignored; but the propionates of the heme of myoglobin are included.

Electrostatic interactions for the groups in these proteins are assigned using the Tanford–Kirkwood method.[2] Charge–charge distances are taken from the Protein Data Bank files. All charge depths are set to 1 Å, and the sphere radius is set to the largest group–group distance found, plus 2. The internal dielectric constant is 2, solvent dielectric constant 80, and ionic strength either 0 or physiologic (150 mM). No Stern layer is used. This combination of parameters yielded rather good agreement with experiment in calculations on the enzyme subtilisin.[25] Self-energies (G_i) are all set to zero. This is likely to make the tests more stringent, since it causes all groups of a given type to have the same intrinsic pK_a, and thus increases the possibility of significant charge–charge correlations.[15] The neglect of desolvation effects may account for what appear to be

unrealistically large, negative values of the ionization energy changes (Table V).

In the cluster calculations, groups whose fractional charge varies by no more than 0.05 when the charges of all other groups are set to their maximal and minimal values are placed in single-group clusters. Cluster method iterations are continued until successive values of θ_i are stable to within 0.000001 protons for each group in the system. The same convergence criterion is used for the iterative mean field method. The reduced site calculations use a cutoff of 0.05 and 0.95 for setting a charge to 0 and 1.0 charge, respectively.

In order for the predominant state and cluster methods to be tested, reliable methods must be used as standards. For small systems, exact or reduced site calculations can serve this purpose. However, these methods are not practical for large systems, because the numbers of states to be enumerated become prohibitively large. For large systems, therefore, the fractional charges computed by the cluster method are compared with the results from the Monte Carlo method of Beroza et al.,[16] using 10,000 steps, as suggested in their work. Because of the possibility that 10,000 steps might not provide adequate convergence, several runs with 50,000 steps were performed. The results agreed well with those of the 10,000 step runs. Tests of the Monte Carlo method on small systems confirmed its validity. This method thus appears to be a very useful standard of comparison for large systems intractable by other methods. However, because the Monte Carlo program as currently constituted does not compute ionization energies, the method is used here only to compute fractional charges. The ionization energies computed by the predominant state method are therefore compared with those from the cluster method. The validity of this approach is discussed below.

RESULTS

Ionization Energy vs. Electrostatic Energy in Conformational Energetics

It has previously been remarked that the energetics of ionizable groups cannot be completely treated in terms of electrostatic interactions.[18] This is a consequence of the fact that electrostatic interactions affect ionization states. The use of a purely electrostatic energy function implicitly assumes that ionization states are fixed.

One striking example of this distinction is that of the desolvation of a highly ionized group, for example a lysine. Consider a lysine having initial pK_a 10.5, in a solvent at pH 7.0. If the side chain is desolvated, with some energy penalty (self-energy) $G_i > 0$, the ionization energy relative to the fully solvated state is

$$\Delta G_{ion} = -RT \ln \frac{1+K_a^{-1}[H^+]e^{-\beta G_i}}{1+K_a^{-1}[H^+]}. \quad (31)$$

In the limiting case of G_i large and positive (group highly desolvated), this expression becomes $2.303RT(pK_a - pH)$, where pK_a is that of the unperturbed group. The reason that a limiting value is reached is that the group becomes neutral, and the limiting value is the energy required to neutralize the group. In the present case, this limiting value proves to be only 4.8 kcal/mol. This represents the maximal pH-dependent component of the desolvation energy of a lysine side chain. To it may be added a "dipole" electrostatic energy (see Methods) for the desolvation of a neutral lysine side chain. This energy is approximated by the work of transferring a neutral primary amine from water to the gas phase, about 4.5 kcal/mol.[26] Thus the net desolvation energy of a lysine side chain at pH 7.0 is only about (4.8 + 4.5 =) 9.3 kcal/mol. This value agrees very well with the "hydration potential" of 9.5 kcal/mol for *n*-butylamine at pH 7.0, as measured by Wolfenden et al.[27] It is quite different from the values of 60–70 kcal/mol used in various molecular energy functions which include solvation effects,[11–13] which are based on the assumption that the group is fixed in its ionized state. Similar corrections will apply for the other ionizable side chains. It is worth noting that simply scaling down the solvation energy parameters for ionizable groups would not be a physically realistic correction for this problem. This is because the forces associated with an ionizable group are purely electrostatic, as derived in Methods. Thus, the forces associated with desolvating, say, an ammonium group will in fact be those associated with the ionized form *until the group becomes neutral,* at which point the forces will suddenly drop in magnitude.

Recent experiments confirm the significance of these theoretical observations. Stites et al.[28] have created a stable mutant of Staphylococcal nuclease containing a completely buried lysine side chain surrounded by nonpolar atoms. The pH dependence of the free energy of unfolding suggests that this group is neutral at physiological pH, with a pK_a of about 6.4. Similarly, Varadarajan et al.[29] have generated a stable mutant of human myoglobin with a buried glutamic acid side chain which appears to be neutral at physiologic pH, with a pK_a of about 8.9. It is improbable that an energy function which does not allow such groups to become neutral would predict these proteins to be stable.

Another way of viewing this situation is that it is rather easy to neutralize an amine or carboxylic acid, at physiologic pH.[30] This implies that carefully assessing the ionization state of a protein before performing a molecular dynamics calculation, for instance, may be more important than is generally recognized.

This discussion has so far focused on the consequences of desolvating an ionizable group. However, similar considerations apply for interactions among ionizable groups. Thus, the common practice of fix-

ing the charge of each ionizable group at an integer value and using a purely electrostatic energy function can lead to errors, even when solvation energies are not included. A sense of the magnitude of these errors can be obtained by comparing purely electrostatic energies with correctly calculated ionization energies for model systems, as now described.

Ten different sets of 10 ionizable groups are randomly generated. The distribution is such that one-half of groups will be acids with pK_as of 4.0, one-quarter will be bases with pK_as of 6.0, and one-quarter will be bases with pK_as of 10.0. Self energies (G_i) are all set to zero, as is the practice in many conformational studies. For each set of ionizable groups thus created, 10 random "conformations" are generated. The groups are scattered randomly over the surface of a sphere of radius 9 Å, with the restriction that no two groups are permitted to lie closer to each other than 3 Å. The groups are then taken to be embedded in a concentric low dielectric sphere of radius 10 Å, which is immersed in high dielectric solvent containing an electrolyte. The internal dielectric constant is set at 2, the external dielectric constant to 80, and the ionic strength to 150 mM (physiologic). The surface density of groups in these systems is comparable to that of hen egg white lysozyme, which has 32 ionizable groups, and a radius of roughly 17 Å.[31]

This procedure generates 10 different "conformations" for each of 10 model "proteins," each having 10 ionizable groups. The Tanford–Kirkwood equations[2] are then used to calculate charge–charge interactions. Comparison is made between true ionization energies for pH 7.0, and electrostatic energies based upon the assumption that each group is either fully ionized or fully neutral, depending upon the pH. Thus, for pH 7.0 the acids of pK_a 4.0 and bases of pK_a 10.0 are fully ionized, and the bases of pK_a 6.0 are fully neutral. (The ionization energies are relative to those for the same set of ionizable groups in the absence of electrostatic interactions.) For the resulting 100 comparisons, the rms difference between the ionization and the electrostatic energies is 4.8 kcal/mol, with maximum difference of 17.4 kcal/mol. The range of calculated ionization energies is −20.4 to 4.9 kcal/mol, so the rms deviation is about 20% of the overall range. Linear regression analysis yields a correlation coefficient of 0.78, with slope of 0.80, intercept of 2.2 kcal/mol, and standard error of 3.4 kcal/mol. Figure 3 provides a scatter plot of ionization versus electrostatic energies. It is worth noting that the data point with the most favorable ionization energy does not have the most favorable electrostatic energy. In summary, this purely electrostatic model, with charges fixed at their unperturbed values, yields rather poor agreement with correctly calculated ionization energies.

Figure 3 also shows that virtually all the electrostatic energies are more positive than the ionization

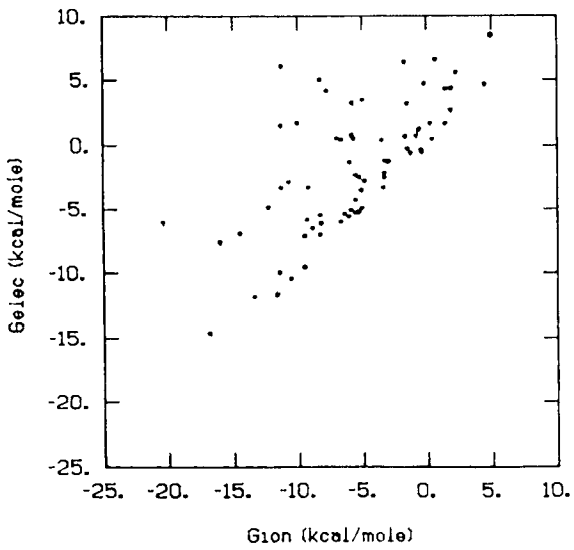

Fig. 3. Scatter plot of ionization energies at pH 7.0 (relative to values for the same groups in unperturbed state at pH 7.0) versus electrostatic energies calculated using the fixed charge assumption (see text).

energies. The explanation for this lies in the fact that the electrostatic energies are calculated under the assumption that ionization states are fixed. In the full ionization calculations, unfavorable interactions are at least partially "relieved" by changes in ionization state, as in the example of the desolvated lysine, above. In the present calculations, most of these changes in ionization state are induced ionizations of histidine residues, which the purely electrostatic calculations treat as neutral. Occasional negative errors may result from forcing groups to have integer charges.

Recent experimental results emphasize the importance of changes in ionization state on protein stability, for example.[32,33] In general, the size of the errors in theoretical energy calculations introduced by neglecting ionization changes will depend upon the system under study. A system with a number of groups having pK_as close to the pH will be more subject to errors, because the fixed charge approximation will lead to more errors in charge state. Errors will also tend to be larger in systems having strong interactions. It is perhaps worth noting that the calculated strengths of electrostatic interactions will depend not only upon the system under study, but also upon the model used for calculating electrostatic interactions.

Tests of Numerical Methods

The results of the previous section demonstrate that neglecting the energetics of ionizable groups in energy calculations can lead to significant errors. The present section provides tests of the predominant

TABLE I. Tests on Small Model Systems*

Coupling	rms	Max	Corr	Slope	Intcpt	SE
(a)						
"Strong"	0.24	1.2	1.00	0.99	0.11	0.22
"Moderate"	0.097	0.52	1.00	1.00	0.035	0.093
(b)						
"Strong"	0.077	0.31	1.00	1.00	0.048	0.057
"Moderate"	0.024	0.14	1.00	1.00	0.014	0.021
(c)						
"Strong"	0.022	0.17	1.00	1.01	0.00	0.021
"Moderate"	0.006	0.062	1.00	0.00	1.00	0.005

*Tests of predominant state and cluster methods for 25 randomly generated systems of 10 groups each, with "strong" or "moderate" coupling, as defined in text. (a) Error analysis of ionization energies of predominant state method. (b) Error analysis of ionization energies of cluster method, with maximum cluster size of 5. (c) Error analysis of fractional charges of cluster method. rms, root-mean-square error; Max, maximum absolute error; Corr, correlation coefficient; Slope, slope of linear regression fit; Intcpt, y-intercept of linear regression fit; SE, standard error of linear regression fit. Charges in proton charges, energies in kcal/mol.

state and cluster methods, which may be helpful in providing rapid evaluations of ionization energies.

Tests on small model systems

The predominant state method is tested first on sets of 10 ionizable groups. The "strong coupling" case of Bashford and Karplus[15] is examined first, as this represents a relatively difficult challenge, where charge–charge correlations will be large. The pK_as of the groups are randomly and uniformly distributed in the range 4.0 to 10.0, and the groups are randomly assigned as acidic or basic with equal probability. All self energies (G_i) are set to zero, and the interaction energies (G_{ij}) are randomly and uniformly distributed in the range of 0.69–2.1 kcal/mol. Twenty-five sets of 10 groups are generated, and the exact ionization energies of each are compared with the results of the predominant state method, implemented as described in Methods. The pH range is 0 to 12, in steps of 0.5, so a total of 2,500 comparisons are made.

As shown in Table I, part (a), the agreement is quite good, with an rms deviation of the approximate from the exact ionization energies of 0.24 kcal/mol. Linear regression analysis also shows an excellent correlation and a small standard error. Although the maximum error is 1.2 kcal/mol, this should be compared with the large overall range of ionization energies of −7 to 33 kcal/mol, with a mean of 6.2 kcal/mol. The "moderate coupling" test case of Bashford and Karplus[15] differs from the "strong coupling" case only in using a range of G_{ij} values from 0.138 to 1.10 kcal/mol. It is therefore a less stringent test. Results of approximate versus exact ionization energies are given in Table I, part (b). As expected, the agreement is better than that obtained for the "strong" coupling case. The maximum error has fallen from 1.2 to 0.52 kcal/mol. Taken together, these tests show that, for a large, challenging set of test cases, the predominant state method yields remarkably accurate ionization energies.

In order to confirm that these test cases are not trivially easy, it is of interest to evaluate the performance of the purely electrostatic, fixed charge approach examined in the previous section. The ionization state of each group is taken to be either −1, 0, or 1 depending upon its intrinsic pK_a relative to the pH, and upon whether it is an acid or base. The purely electrostatic energy of all 25 "strong coupling" systems is then compared with the correct ionization energy. For pH 7.0, the rms error is 4.6 kcal/mol, with a maximum error of 13.2 kcal/mol. For the "moderate coupling" systems, the rms error is 1.7 kcal/mol, maximum error 5.1 kcal/mol. These rather large errors verify that the success of the predominant state does not result from the use of easy test cases.

The cluster method is tested on the same "strong coupling" and "moderate coupling" sets of 10 groups. Clusters are limited to a maximum size of 5 groups by adjusting the cutoff energy (G_c), as described in Methods. (Small clusters are used here simply to force a test of the method in systems which each contain only 10 groups.) The results for ionization energies are shown in Table I, part (b). Not surprisingly, the results are superior to those obtained with the simpler predominant state method. The accuracy of fractional charges calculated by the cluster method is also assessed for these systems, by comparison with the exact results. The results are in Table I, part (c). Quite good overall results are obtained, although the maximal error of 0.17 for the "strong" coupling case is not as good as might be hoped. However, it should be noted that these test systems are particularly challenging for the cluster method, because the groups simply do not tend to form natural clusters. This is a consequence of the fact that any group has a good chance of interacting with any other one, so the entire systems tends to

TABLE II. Illustration of Cluster Method*

pH	Exact		Cluster		Mean field		Predom
	θ	G_{ion}	θ	G_{ion}	θ	G_{ion}	G_{ion}
0	0.886	50.5	0.888	50.5	0.921	50.8	50.1
1	0.708	46.6	0.708	46.7	0.814	48.3	46.8
2	0.558	39.4	0.557	39.4	0.684	43.4	38.9
3	0.506	30.2	0.506	30.2	0.545	36.0	27.6
4	0.484	20.6	0.484	20.6	0.405	25.9	18.3
5	0.399	10.7	0.401	10.7	0.268	13.8	12.3
6	0.225	2.44	0.225	2.45	0.144	3.43	1.6
7	0.068	.119	0.066	0.127	0.051	0.230	0.01

*Comparison of exact, cluster, mean field, and predominant state methods for model system consisting of 7 clusters of 2 groups each (see text for details). θ, average charge of a group as a function of pH (one value given because all groups equivalent); G_{ion}, ionization energy of the system, relative to that of the same set of groups at the same pH but without electrostatic interactions. Units as in Table I.

form one large cluster. Thus, the cutoff energies required to separate each set of ten groups into clusters of five or fewer (see Methods) ranged up to the largest interaction energy allowed in the sets (1.1 and 2.1 kcal/mol for the "moderate" and "strong" cases, respectively). The situation in real proteins will usually be more conducive to the separation of groups into weakly interacting clusters, because groups which are far apart interact weakly, while in these random test cases there is no such geometric constraint.

The basic validity of the cluster method is better illustrated in the following somewhat artificial test case, consisting of 14 basic sites, each having a pK_a of 6.0. The groups are divided into 7 clusters of two groups each. The interaction energy for each pair of groups is 5 kcal/mol. The interaction energy of each group with each group in a different cluster is 0.2 kcal/mol. An accurate treatment of this system in the pH range from 0 to 6 will clearly require a method which deals with intracluster charge–charge correlations, because each cluster consists of two strongly interacting groups with the same pK_a.[15] In addition, the reduced site method is of no value in this pH range, because all the groups are partially ionized. Table II presents a comparison of charges and ionization energies calculated exactly, with those calculated by means of the cluster method, the mean field method of Tanford and Roxby, and the predominant state method (energies only). The cluster method yields charges and energies that agree with the exact results to high precision. This observation supports the method's basic premise that as long as strong interactions are dealt with in detail, weak interactions can be treated approximately. Although it might be supposed that the cluster method succeeds here merely because the interactions between clusters are negligibly weak, this is not the case. Thus, when the exact calculation is repeated at pH 0, for example, with all intercluster interactions set to zero, the average charge of a group rises from 0.886 to 0.996, and the ionization energy of the system falls from 50.5 to 35.0 kcal/mol. This confirms that the intercluster interactions are not negligible.

Perhaps surprisingly, the predominant state energies are more accurate than those of the mean field method, although the method is simpler and faster. This appears to be a consequence of the fact that the predominant state method can yield states in which one group of each cluster is charged and the other neutral. In contrast, the mean field method enforces symmetry and neglects correlations, and therefore always treats the groups as simultaneously charged. The repulsive interactions then lead to excessively positive ionization energies, as shown in Table II.

Lysozyme

The case of hen egg white lysozyme at 0.1 M ionic strength has been treated by Bashford and Karplus[8] and Beroza et al.,[16] and serves as a useful test case. However, it has been noted that lysozyme is a particularly tractable case because the charge–charge interactions are relatively weak (D. Bashford, personal communication). In order to create a more challenging test, the present calculations use the intrinsic pK_as calculated by Bashford and Karplus using the linearized Poisson–Boltzmann equation,[8] but the site–site interactions calculated by them have all been multiplied by a factor of three. Despite this modification, the reduced site method can still be used as a basis for comparison.

Table III compares ionization energies and fractional charges for this system, calculated using the reduced site, cluster, predominant state, and Monte Carlo methods. The maximum cluster size allowed is somewhat arbitrarily set to 10 here. (The issue of cluster size is examined below.) Computation times are also presented. The reduced site method is used as a standard, although as noted below, this may not be entirely appropriate. (The expression for ionization energy in a system treated by the reduced site method is derived in Appendix B.) The results may

TABLE III. Tests on Lysozyme*

	RS	Cluster			Predom	Monte Carlo		CPU time (sec)			
pH	G_{ion}	G_{ion}	rms	Max	G_{ion}	rms	Max	RS	CL	PD	MC
0	23.4	23.4	0.010	0.041	23.4	0.011	0.044	0.01	0.05	0.02	69
1	18.7	18.7	0.005	0.022	18.6	0.005	0.022	0.02	0.18	0.02	69
2	12.0	12.1	0.001	0.003	11.8	0.003	0.008	0.04	0.27	0.01	69
3	3.29	3.31	0.000	0.001	3.32	0.004	0.012	0.09	0.59	0.02	69
4	−4.85	−4.85	0.001	0.003	−4.86	0.002	0.011	0.23	0.06	0.01	69
5	−7.76	−7.76	0.000	0.002	−7.77	0.002	0.009	0.22	0.05	0.01	69
6	−7.96	−7.96	0.000	0.001	−7.96	0.001	0.004	5.62	0.20	0.01	69
7	−7.54	−7.53	0.001	0.002	−7.53	0.001	0.002	1.18	0.09	0.01	69
8	−6.97	−6.96	0.000	0.000	−6.96	0.002	0.006	1.13	0.70	0.01	69
9	−6.74	−6.73	0.000	0.000	−6.85	0.003	0.007	0.23	0.59	0.01	69
10	−5.84	−5.83	0.000	0.000	−5.66	0.003	0.014	0.50	0.59	0.01	69

*Comparison of reduced site (RS), cluster (CL), predominant state (PD), and Monte Carlo (MC) methods for lysozyme with all group–group interactions multiplied by 3 but unchanged intrinsic pK_as (see text). G_{ion}, ionization energy relative to that of the same set of groups with no electrostatic perturbations (model compound pK_as); rms, root-mean-square difference of average charge ($z(i)\theta_i$) of all groups, relative to reduced site results; Max, maximum absolute error in charge for any group. Approximate CPU times are given in seconds for each method. Units as in Table I.

be summarized as follows. All the ionization energies agree well with each other. This is particularly remarkable for the predominant state method, given its simplicity and speed. All the fractional charges are generally in good agreement with the reduced site results. However, for both the Monte Carlo and cluster methods, the agreement is worse at pH 0 and 1. It turns out that the Monte Carlo and cluster methods agree with each other better than they do with the reduced site method. For example, at pH 0, the rms deviation and maximum deviation for these two methods are 0.003 and 0.013 protons, respectively; at pH 1, the values are 0.002 and 0.006. It is probable that the Monte Carlo and cluster methods are actually slightly more accurate than the reduced site method in this case. In the cluster calculations, the largest cutoff energy (G_c) required to generate clusters of 10 groups or fewer was 0.4 kcal/mol.

The computation times are extremely short for the predominant state method, with a maximum of 0.02 sec. At some pHs, the reduced site method is more rapid than the cluster method, because many groups can be fixed, and because no iterations are involved. However, at other pHs, the reduced site method is slower, because only a few groups can be treated as fixed. The ability of the cluster method to separate the nonfixed groups into clusters of manageable size gives it a speed advantage under these circumstances. The Monte Carlo method is uniformly slow, with computation times which are independent of the pH.

Other proteins

For all the other protein systems examined, the Tanford–Kirkwood formulas[2] are used to estimate charge–charge interaction energies, using parameters which have been shown to yield fairly good agreement with a set of experimental data (see Methods). The advantage of using actual protein coordinates, rather than random distributions of ionizable groups on the surfaces of spheres, is that there is some validity to the site distribution, and the value of the cluster method will depend to some extent on just how ionizable groups are distributed. The proteins examined are poplar plastocyanin, with 27 ionizable groups; sperm whale deoxymyoglobin, with 62 ionizable groups; rhodanese, with 93 ionizable groups; and pig heart citrate synthase, with 123 ionizable groups.

Table IV compares the results of fractional charge calculations using the cluster method with results from 10,000-step Monte Carlo calculations. Maximum cluster sizes of 1, 2, 5, and 10 are examined. Each reported value represents combined results for pHs 2, 4, 6, 8, and 10. Also reported are the average time for a calculation at a single pH by the cluster method, and the largest value of G_c (see Methods) required to separate the groups into clusters within the allowed size limit.

As expected the results improve with increasing cluster size. Although the rms results are not bad for clusters of one group each (i.e., for the pure mean field method), the maximum errors are quite large in this case. However, the results are excellent when clusters of up to 10 groups are allowed. Here, the overall maximum error falls to 0.072 protons, for rhodanese at zero ionic strength, and the rms errors are consistently less than 0.01 proton. The computations are still quite rapid with 10-group clusters. Average CPU times range from 0.7 to 15 sec. This may be compared with the times required for a Monte Carlo run at a single pH, which range from 47 sec for plastocyanin to 850 sec for citrate synthase.

It is perhaps surprising that the computation times fall as the allowed cluster size increases from 1 to 5. This is a consequence of the large number of itera-

TABLE IV. Tests of Cluster Method for Four Proteins*

Max clust	Physiologic ionic str. rms	Max	CPU(s)	G_c(max)	Zero ionic str. rms	Max	CPU(s)	G_c(max)
				Plastocyanin				
1	0.087	0.38	0.28	4.8	0.12	0.46	0.6	7.4
2	0.018	0.12	0.18	2.7	0.037	0.31	0.2	3.9
5	0.005	0.025	0.08	1.6	0.004	0.019	0.14	2.0
10	0.005	0.025	0.65	0.4	0.004	0.015	1.3	0.9
				Sperm whale deoxymyoglobin				
1	0.050	0.87	0.64	13.0	0.050	0.88	1.3	13.6
2	0.004	0.041	0.48	4.5	0.006	0.048	0.88	5.0
5	0.004	0.041	0.44	2.2	0.003	0.032	0.68	2.9
10	0.003	0.036	2.2	1.7	0.003	0.034	4.7	2.2
				Rhodanese				
1	0.058	0.71	2.3	8.5	0.055	0.51	3.5	9.1
2	0.012	0.14	1.1	5.9	0.014	0.12	2.6	6.5
5	0.005	0.039	0.84	2.7	0.012	0.12	2.1	3.2
10	0.004	0.028	3.1	1.0	0.007	0.072	10.0	1.8
				Citrate synthase				
1	0.15	0.86	3.6	15.9	0.032	0.33	8.2	16.5
2	0.022	0.31	2.4	8.3	0.015	0.19	5.6	8.9
5	0.004	0.051	1.8	3.7	0.006	0.053	4.1	4.3
10	0.003	0.037	5.2	2.4	0.006	0.050	15.2	2.9

*Error analysis of fractional charges of cluster method relative to Monte Carlo method for four proteins. Max clust, largest cluster size allowed; G_c(max), largest intercluster interaction energy required to keep clusters within allowed limit. Definitions otherwise as in other tables.

tions required to achieve convergence when the pure mean field method is used, and the relatively low computational cost associated with small clusters. When clusters of 10 groups are allowed, the computational times increase, and in fact the times increase rapidly when still larger clusters are permitted. For example, a calculation on citrate synthase at zero ionic strength requires an average of 150 sec for each pH when clusters of up to 14 groups are permitted.

In most cases, very good results are obtained with maximum cluster sizes of only 5 groups, and the computation times are quite low. It would seem that allowing very large clusters will be unnecessary in many cases, and that when computer time is at a premium, the use of a small cluster size may well be justifiable. In the present calculations, the computer times required for the runs with a maximum cluster size of 5 ranges between 0.08 sec for plastocyanin to 4 sec for citrate synthase, and the largest error is 0.12 protons.

It is also of interest to analyze the results in terms of the residual inter-cluster energies, [G_c(max)], because the results can be expected to become more accurate as these values fall. In fact, the correlation between these energies and the errors is surprisingly poor. Remarkably good results are frequently obtained even with fairly large values of G_c(max). For example, for plastocyanin with a maximum cluster size of 5, the maximum error in a fractional charge is only 0.019, even though at least one intercluster interaction of ~ 2 kcal/mol exists. This contrasts with the results for the "strong" and "moderate" coupling tests of Bashford and Karplus (see above), where a fractional charge error of 0.17 was found when intercluster interactions were still ~ 2.1 kcal/mol. The fact that better results are obtained for proteins probably results from the fact that the groups form clusters more readily, because only neighboring groups interact strongly. As a consequence, there are fewer of these large residual intercluster interactions. It should also be noted that treating a strong interaction by the mean field approximation does not necessarily result in a large error. For an error to be large, it must also be the case that the groups have similar pK_as, making charge–charge correlations significant.

The fact that the cluster method yields excellent values for fractional charges in these protein systems strongly suggests that the ionization energies it yields are also excellent. In fact, they are probably more accurate than the fractional charges, because it is in general easier to compute ionization energies than charges. For example, the predominant state method yields rather accurate energies in the systems so far analyzed, but not particularly accurate charges. Also, the cluster method yields ionization energies accurate to within 0.31 kcal/mol for the "strong" coupling case (see above), where the largest

TABLE V. Tests of Predominant State Method for Four Proteins*

	Plastocyanin		Myoglobin		Rhodanese		Citrate synthase	
pH	CL	PD	CL	PD	CL	PD	CL	PD
2	−9.65	−9.73	−22.6	−23.4	9.50	9.45	12.2	12.8
4	−27.7	−27.5	−90.7	−91.1	−73.9	−73.9	−94.8	−96.0
6	−24.8	−24.2	−118.6	−118.7	−101.1	−101.3	−130.4	−131.7
8	−14.8	−15.2	−119.8	−119.9	−101.1	−100.4	−128.4	−127.6
10	−4.21	−4.20	−113.1	−113.1	−93.1	−92.5	−120.0	−120.4
2	−8.12	−8.16	−25.5	−26.1	−5.29	−5.11	−15.3	−16.0
4	−24.2	−24.5	−83.7	−84.3	−66.2	−66.3	−90.3	−90.4
6	−22.9	−23.5	−107.0	−107.1	−86.1	−86.2	−112.1	−112.5
8	−15.5	−16.1	−106.5	−106.6	−85.6	−84.3	−108.8	−108.6
10	−7.87	−7.92	−100.4	−100.5	−81.3	−81.4	−107.1	−107.7

*Comparison of ionization energies of predominant state method (PD) with results of cluster method (CL) using maximum cluster size of 10. Top half of table: physiologic ionic strength. Bottom half of table: zero ionic strength. Energies in kcal/mol.

error in fractional charge is 0.17, which is more than twice the largest error found in the protein systems examined here.

The ionization energies provided by the cluster method may therefore be used as standards for testing ionization energies calculated by the predominant state method. (As noted above, the Monte Carlo method does not permit the calculation of ionization energies.) Table V compares ionization energies computed by the predominant state method with those computed by the cluster method using a maximum cluster size of 10, for the four protein models, for pHs 2, 4, 6, 8, and 10. The agreement is excellent. The largest difference between the two methods is 1.3 kcal/mol, for rhodanese at zero ionic strength at pH 8, and for citrate synthase at physiologic ionic strength at pH 6. The percentage errors involved are less than 2%. Moreover, the computation times are very small, ranging from 0.01 sec for plastocyanin to 0.34 sec for citrate synthase. This speed, in unoptimized code, raises the possibility of incorporating ionization energies in conformational energy calculations with manageable computational costs.

DISCUSSION

As demonstrated in Results, the neglect of ionization equilibria can lead to sizable error in conformational energies. Although not specifically analyzed here, it is evident that a similar problem exists for forces involving ionizable groups. The degree to which such errors need to be viewed as a problem in current methodologies depends upon, first, the specific systems examined, since errors are likely to be system-dependent; and second, the presumed precision of the energy functions in use. Thus, to the extent that the various energy components in an energy function are viewed as approximate, it may not be worth adding the complications of ionization equilibria. However, given that the errors found in the model computations here range up to about 15 kcal/mol, it is probable that reliable predictions will ultimately require that ionization equilibria be dealt with.

It is perhaps worth reiterating that the energy errors incurred by fixing the ionization states of groups at their unperturbed values will almost always be in the positive direction, because changes in ionization state are always such as to reduce the energy of the system. Therefore computations which neglect ionization changes will tend to underestimate the stabilization afforded by electrostatic interactions of ionizable groups. A particularly striking example is that of the desolvation of a lysine side chain. Although desolvation of a fixed ionized primary amine costs 60–70 kcal/mol, the energy cost of desolvating the same group in equilibrium with a bath of pH 7.0 is actually less than 10 kcal/mol, as confirmed experimentally. The cost will, of course, vary with pH. Note also that a neutral buried lysine in a medium of pH 7.0 does destabilize the system. That is, the group does not have to be ionized for destabilization to occur, and a purely electrostatic model which treats the group as neutral will actually underestimate the cost of desolvation.

In cases where ionization effects are important, an array of methods is now available. Exact calculations and the reduced site method[15] are convenient and fast for small systems; the powerful Monte Carlo method[16] will likely remain a standard for the largest systems; and for systems of intermediate size, the methods described here should be of considerable utility. The predominant state method is the simplest and fastest. CPU times in the systems examined here ranged from less than 0.01 sec to a maximum of 0.34 sec. Greater efficiency could be obtained with optimization, and perhaps vectorization, of the computer program. In addition, there may well be more efficient techniques for selecting highly occupied states than the one proposed here. It may thus be possible in the near future to include a predominant state method in the energy computations of a conformational search routine.

On the other hand, because the predominant state method does not yield particularly accurate fractional charges, it is not well adapted for the calculation of forces. When force calculations are needed, such as in energy minimizations, the cluster method will be preferable. The fractional charges and charge–charge correlations yielded by this method can be substituted into the force expressions derived in Methods, keeping in mind that the charge states of groups in different clusters must be treated as uncorrelated ($\langle x(i)x(j)\rangle = \langle x(i)\rangle\langle x(j)\rangle$). As currently implemented, this approach is probably too slow to be used at every step of an energy minimization, but it may well be feasible to rerun it only every N steps to update $\langle x(i)\rangle\langle x(j)\rangle$, and $\langle x(i)x(j)\rangle$, while updating terms of the form $\partial G_i/\partial y$, $\partial G_j/\partial y$, $\partial G_{ij}/\partial y$ [see Eq. (18)] at every step. In addition, it should be possible to optimize the program considerably. For example, there may be alternative criteria for separating groups into clusters which will provide accurate results with smaller clusters. It should also be possible to devise a more rapid algorithm for forcing this iterative method to converge. As it stands, in any case, the cluster method should be helpful in titration and stability calculations on structures of moderately large size. In such applications, it offers a convenient means of rapidly computing pH-dependent quantities.

CONCLUSIONS

1. A purely electrostatic treatment of energies and forces associated with ionizable groups can be quite inaccurate. For example, the cost of desolvating a lysine at pH 7.0 is actually about 9 kcal/mol, rather than the 60–70 kcal/mol commonly employed in solvation energy computations. In general, the neglect of ionization changes tends to lead to overestimates of free energies.

2. The free energy of a single highly occupied ionization state is a good approximation to the free energy of the system as a whole. Estimates of ionization energies by this approach are accurate and computationally rapid, even in large systems.

3. A second new method groups strongly interacting sites into clusters, and treats intracluster interactions exactly, while using the mean field approximation for cluster–cluster interactions. This represents a computationally efficient approach to computing accurate charges and energies in large systems.

ACKNOWLEDGMENTS

This work was supported by the NIH and the Robert A. Welch Foundation. M. K. Gilson is a Howard Hughes Medical Institute Physician Research Fellow. I am grateful to P. Beroza and co-authors for kindly furnishing the source code for the Monte Carlo titration program[16]; to D. Bashford and M. Karplus for making available the results of electrostatic calculations on lysozyme; to M. Zacharias and H.S.R. Gilson for their thoughtful readings of the manuscript; and to Professor J.A. McCammon for support and suggestions.

REFERENCES

1. Mehler, E.L., Eichele, G. Electrostatic effects in water-accessible regions of proteins. Biochemistry 23:3887–3891, 1984.
2. Tanford, C., Kirkwood, J.G. Theory of protein titration curves. I. General equations for impenetrable spheres. J. Am. Chem. Soc. 79:5333–5339, 1957.
3. Shire, S.J., Hanania, G.I.H., Gurd, F.R.N. Electrostatic effects in myoglobin. Hydrogen ion equilibria in sperm whale ferrimyoglobin Biochemistry 13:2967–2974, 1974.
4. Russell, S.T., Warshel, A. Calculations of electrostatic energies in proteins. The energetics of ionized groups in bovine pancreatic trypsin inhibitor. J. Mol. Biol. 185:389–404, 1985.
5. Zauhar, R.J., Morgan, R.S. A new method of computing the macromolecular electric potential J. Mol. Biol. 186: 815–820, 1985.
6. Warwicker, J., Watson, H.C. Calculation of electric potential in the active site cleft due to alpha-helix dipoles J. Mol. Biol. 157:671–679, 1982.
7. Gilson, M.K., Sharp, K.A., Honig, B.H. Calculating electrostatic interactions in bio-molecules: method and error assessment. J. Comput. Chem. 9:327–335, 1987.
8. Bashford, D., Karplus, M. pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. Biochemistry 29:10219–10225, 1990.
9. Davis, M.E., Madura, J.D., Luty, B.A., McCammon, J.A. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian dynamics program. Comput. Phys. Commun. 62:187–197, 1991.
10. Juffer, A.H., Botta, E.F.F., van Keulen, B.A.M., van der Ploeg, A., Berendsen, H.J.C. The electric potential of a macromolecule in a solvent: A fundamental approach. J. Comput. Phys. 97:144–170, 1991.
11. Vila, J., Williams, R.L., Vasquez, M., Scheraga, H.A. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. Proteins 10:199–218, 1991.
12. Kang, Y.K., Nemethy, G., Scheraga, H.A. Free energies of hydration of solute molecules. 2. Application of the hydration shell model to charged organic molecules. J. Phys. Chem. 91:4109–4120, 1987.
13. Cramer, C.J., Truhlar, D.G. An SCF solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. Science 256:213–216, 1992.
14. Tanford, C., Roxby, R. Interpretation of protein titration curves. Application to lysozyme. Biochemistry 11:2192–2198, 1972.
15. Bashford, D., Karplus, M. Multiple-site titration curves of proteins: An analysis of exact and approximate methods for their calculation. J. Phys. Chem. 95:9556–9561, 1991.
16. Beroza, P., Fredkin, D.R., Okamura, M.Y., Feher, G. Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of Rhodobacter sphaeroides. Proc. Natl. Acad. Sci. U.S.A. 88:5804–5808, 1991.
17. Schellman, J.A. Macromolecular binding. Biopolymers 14: 999–1018, 1975.
18. Tanford, C. Protein Denaturation Part C. Theoretical models for the mechanisms of denaturation. Adv. Prot. Chem. 24:1–95, 1970.
19. Yang, A.-S., Honig, B. Electrostatic effects on protein stability. Curr. Opinion Struct. Biol. 2:40–45, 1991.
20. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.J. The protein data bank: A computer based archival file for molecular structures. J. Mol. Biol. 112:535–542, 1977.
21. Guss, J.M., Harrowell, P.R., Murata, M., Norris, V.A., Freeman, H.C. Crystal structure analyses of reduced Cu(I) poplar plastocyanin at six pH values. J. Mol. Biol. 192:361, 1986.

22. Phillips, S.E.V. Structure refinement of oxymyoglobin at 1.6 Angstrom resolution. J. Mol. Biol. 142:531–554, 1980.
23. Ploegman, J.H., Drent, G., Kalk, K.H., Hol, W.J.G. Structure of bovine liver rhodanese. I. Structure determination at 2.5 Angstroms resolution and a comparison of the conformation and sequence of its two domains. J. Mol. Biol. 123:557–594, 1978.
24. Weigand, G., Remington, S., Deisenhofer, J., Huber, R. Crystal structure analysis and molecular model of a complex of citrate synthase with oxaloacetate and S-acetonylcoenzyme A. J. Mol. Biol. 174:205, 1984.
25. Gilson, M.K., Honig, B.H. Energetics of charge-charge interactions in proteins. Proteins 3:32–52, 1988.
26. Ben-Naim, A., Marcus, Y. Solvation thermodynamics of nonionic solutes. J. Chem. Phys. 81:2016–2027, 1984.
27. Wolfenden, R., Andersson, L., Cullis, P.M., Southgate, C.C.B. Affinities of amino acid side chains for water. Biochemistry 20:849–855, 1981.
28. Stites, W.E., Gittis, A.G., Lattman, E.E., Shortle, D. In a staphylococcal nuclease mutant the side chain of a lysine replacing Val 66 is fully buried in the hydrophobic core. J. Mol. Biol. 221:7–14, 1991.
29. Varadarajan, R., Lambright, D.G., Boxer, S.G. Electrostatic interactions in wild-type and mutant recombinant human myoglobins. Biochemistry 28:3771–3781, 1989.
30. Honig, B.H., Hubbell, W.L. Stability of "salt bridges" in membrane proteins. Proc. Natl. Acad. Sci. U.S.A. 81:5412–5416, 1984.
31. Imoto, T. Electrostatic free energy of lysozyme. Biophys. J. 44:293–298, 1983.
32. Pace, C.N., Laurents, D.V., Erickson, R.E. Urea denaturation of barnase: pH dependence and characterization of the unfolded state. Biochemistry 31:2728–2734, 1992.
33. Sancho, J., Serrano, L., Fersht, A.R. Histidine residues at the N- and C-termini of alpha-helices: Perturbed pKas and protein stability. Biochemistry 31:2253–2258, 1992.

APPENDIX A: ALTERNATIVE DERIVATION OF CLUSTER METHOD ENERGY

Expressions for fractional charges and ionization energy for the cluster method are derived here using an alternative approach to that in the Methods section. For simplicity, the derivation is performed for a small system consisting of one cluster of two ionizable groups (1 and 2), interacting with a second cluster of one group (3). Equation (9) of Bashford and Karplus[15] becomes

$$\begin{aligned} G_{\text{ion}} = \min\ & [\beta^{-1}\rho(000)\ln\rho(000)] + \rho(100)\,[b_1 + \beta^{-1}\ln\rho(100)] \\ & + \rho(010)\,[b_2 + \beta^{-1}\ln\rho(010)] + \rho(001)\,[b_3 + \beta^{-1}\ln\rho(001)] \\ & + \rho(110)\,[b_1 + b_2 + G_{12} + \beta^{-1}\ln\rho(110)] \\ & + \rho(101)\,[b_1 + b_3 + G_{13} + \beta^{-1}\ln\rho(101)] \\ & + \rho(011)\,[b_2 + b_3 + G_{23} + \beta^{-1}\ln\rho(011)] \\ & + \rho(111)\,[b_1 + b_2 + b_3 + G_{12} + G_{13} + G_{23} + \beta^{-1}\ln\rho(111)] \end{aligned} \tag{32}$$

where the argument of ρ gives the ionization states of the three groups, and the minimization is subject to the requirement that the ρ values sum to 1.

In the cluster method, the ionization states of the two clusters are taken to be uncorrelated. For example, ρ(000) = ρ(00)ρ(0), where the number of arguments for ρ is two for cluster I and one for cluster II. By the use of such expressions for all the values of ρ, G_{ion} can be rewritten

$$\begin{aligned} G_{\text{ion}} = \min_{\rho_{\text{I}},\rho_{\text{II}}}\{ & \rho(00)\beta^{-1}\ln\rho(00) + \rho(10)\,[b_1 + \beta^{-1}\ln\rho(10)] \\ & + \rho(01)\,[b_2 + \beta^{-1}\ln\rho(01)] \\ & + \rho(11)\,[b_1 + b_2 + G_{12} + \beta^{-1}\ln\rho(11)] \\ & + \rho(0)\beta^{-1}\ln\rho(0) + \rho(1)\beta^{-1}\ln\rho(1) \\ & + \theta_1\theta_3 G_{13} + \theta_2\theta_3 G_{23}\} \end{aligned} \tag{33}$$

where it is now required that the probability distribution for each cluster sum to 1. The combination of Eq. (33) and the constraints

$$\begin{aligned} \rho(00) + \rho(10) + \rho(01) + \rho(11) - 1 &= 0 \\ \rho(0) + \rho(1) - 1 &= 0 \end{aligned} \tag{34}$$

may be solved by the method of Lagrangian multipliers to yield

$$\begin{aligned} \rho(00) &= [1 + e^{-\beta(b_1+\theta_3 G_{13})} + e^{-\beta(b_2+\theta_3 G_{23})} + e^{-\beta(b_1+b_2+G_{12}+\theta_3(G_{13}+G_{23}))}]^{-1} \\ \rho(10) &= e^{-\beta(b_1+\theta_3 G_{13})}\,\rho(00) \\ \rho(01) &= e^{-\beta(b_2+\theta_3 G_{23})}\,\rho(00) \\ \rho(11) &= e^{-\beta(b_1+b_2+G_{12}+\theta_3(G_{13}+G_{23}))}\,\rho(00) \\ \rho(0) &= [1 + e^{-\beta(b_3+\theta_1 G_{13}+\theta_2 G_{23})}]^{-1} \\ \rho(1) &= e^{-\beta(b_3+\theta_1 G_{13}+\theta_2 G_{23})}\,\rho(0). \end{aligned} \tag{35}$$

Because the Bashford and Karplus paper uses a reference state having all groups deprotonated, while the present work defines the reference state having all groups neutral, a slightly different definition of b_i is required:

$$b_i \equiv G_i - 2.303RT\,z_i(\text{p}K_{\text{a}_i} - \text{pH}) \tag{36}$$

Here we use the model compound $\text{p}K_{\text{a}}$, instead of the intrinsic $\text{p}K_{\text{a}}$, so the self-energy G_i is included explicitly in the definition. It is simple to show that

$$\begin{aligned} \rho(00) &= \Sigma_I^{-1} \\ \rho(0) &= \Sigma_{II}^{-1} \end{aligned} \tag{37}$$

where Σ_{I}^{-1} and Σ_{II}^{-1} are the ionization polynomials of the two clusters, as defined in Methods.

Substition of Eqs. (35), (36), and (37) into Eq. (33) yields an energy expression which simplifies to

$$\begin{aligned} G = & -RT\ln\Sigma_I - RT\ln\Sigma_{II} \\ & - \theta_3(\theta_1 G_{13} + \theta_2 G_{23}) \end{aligned} \tag{38}$$

which agrees with Eq. (30) for this three-group system.

APPENDIX B: IONIZATION ENERGY IN THE REDUCED SITE APPROXIMATION

We wish to derive an expression for the ionization energy of a system containing some ionizable groups which are essentially fully ionized, some which are essentially fully neutral, and some which are partially ionized. In the reduced site method, the fully

ionized and fully neutral groups are removed from the enumeration of ionization states. In writing the ionization polynomial for this system, we first note that the fully neutral groups contribute nothing to the ionization energy, because the reference state has all the groups neutral. The ionization states which do contribute to the ionization energy all have a subset of groups fixed fully ionized. The contribution of these groups can be factored out of the ionization polynomial. If there are n_f fixed ionized groups and n nonfixed groups, the ionization energy takes the form

$$G_{\text{ion}} = -RT \ln \left(\prod_{i_f}^{n_f} A_{i_f} e^{-\beta \left[\sum_{i_f}^{n_f} \left(G_{i_f} + \sum_{j_f > i_f}^{n_f} G_{i_f j_f} \right) \right]} \times \left[1 + \sum_{i}^{n} A_i e^{-\beta \left(G_i + \sum_{i_f}^{n_f} G_{i i_f} \right)} + \cdots + \left(\prod_{i}^{n} A_i \right) e^{-\beta \left(\sum_{i}^{n} \left(G_i + \sum_{j>i}^{n} G_{ij} + \sum_{i_f}^{n_f} G_{i i_f} \right) \right)} \right] \right) \tag{39}$$

where i_f and j_f index only the fixed groups, and i and j index only the nonfixed groups. This expression may be rewritten

$$G_{\text{ion}} = 2.303\, RT \sum_{i_f}^{n_f} z_{i_f} (pH - pK_{a_{i_f}}) + \sum_{i_f}^{n_f} \left(G_{i_f} + \sum_{j_f > i_f}^{n_f} G_{i_f j_f} \right) - RT \ln \Sigma^*$$

where Σ^* is the ionization polynomial for the nonfixed groups, where the electrostatic energy of each state includes the interactions of the nonfixed with the fixed groups. The first term here is the ionization energy of the fixed groups in the absence of any interactions; the second term is the electrostatic interaction energy of the fixed groups with each other; and the final term is the ionization energy of the nonfixed groups in the electrostatic field of the fixed groups.