PROTEINS: Structure, Function, and Bioinformatics 63:928–938 (2006)

A Simple Clustering Algorithm Can Be Accurate Enough for Use in Calculations of p*K*s in Macromolecules

Jonathan Myers, Greg Grothaus, Shivaram Narayanan, and Alexey Onufriev*
Department of Computer Science, Virginia Tech, Blacksburg, Virginia

ABSTRACT **Structure and function of macromolecules depend critically on the ionization states of their acidic and basic groups. Most current structure-based theoretical methods that predict p*K* of ionizable groups in macromolecules include, as one of the key steps, a computation of the partition sum (Boltzmann average) over all possible protonation microstates. As the number of these microstates depends exponentially on the number of ionizable groups present in the molecule, direct computation of the sum is not realistically feasible for many typical proteins that may have tens or even hundreds of ionizable groups. We have tested a simple and robust approximate algorithm for computing these partition sums for macromolecules. The method subdivides the interacting sites into independent clusters, based upon the strength of site–site electrostatic interaction. The resulting partition function is factorizable into computationally manageable components. Two variants of the approach are presented and validated on a representative test set of 602 proteins, by comparing the p$K_{1/2}$ values computed by the proposed method with those obtained by the standard Monte Carlo approach used as a reference. With 95% confidence, the relative error introduced by the more accurate of the two methods is less than 0.25 pK units. The algorithms are one to two orders of magnitude faster than the Monte Carlo method, with the typical settings. A graphical representation is introduced that visualizes the clusters of strong site–site interactions in the context of the three-dimensional (3D) structure of the macromolecule, facilitating identification of functionally important clusters of ionizable groups; the approach is exemplified on two proteins, *bacteriorhodopsin* and *myoglobin*. Proteins 2006;63:928–938.** © 2006 Wiley-Liss, Inc.

Key words: titration; protonation; clustering; electrostatics; p*K*

INTRODUCTION

Electrostatic interactions are often a key factor determining properties of biomolecules[1–7] including their biological functions such as catalytic activity,[8,9] ligand binding,[10] complex formation,[11] proton transport,[12] as well as structure and stability.[13–15]

The electrostatic properties of a molecule can change dramatically depending on the ionization (protonation) states of its titratable groups. The protonation of a group depends on the group's type, location within the macromolecule, ionization state of other titratable sites, and the pH and ionic strength of the surrounding solvent. There are suggestions that link such dramatic changes to the associated protein function.[16]

A wide range of theoretical methods exists that predict pK_a and protonation states of ionizable groups,[17–31] to cite just some. Apart from the very early approaches[17,32] which represented the molecule as a low dielectric sphere, and made mostly qualitative predictions, all modern methods use atom-detail information from high resolution Protein Data Bank (PDB) structures; generally, higher resolution data yields more accurate predictions. While these methods vary in the underlying physical models and accuracy, they share one common feature—computational intensity. For most of the approaches, this computation intensity stems, in general, from two sources: the need to accurately estimate charge–charge interactions inside an arbitrarily shaped molecule, and the need to compute the appropriate partition function (statistical sum) over all possible microstates. In this work, we focus on reducing the computational intensity of the second step, and assume that the free energy of each microstate can be calculated by the now standard continuum electrostatics methods described elsewhere.[33] Recent advances in the implicit solvent methodology, in particular the Generalized Born approximation, have significantly reduced the computational costs of estimating charge–charge interactions.[34]

To be specific, consider a molecule with $\mathbf{n}$ titratable sites. Assume, as is typical in the field, that the protonation microstate of any individual site is binary, that is either protonated (1) or de-protonated (0). A protonation state of the entire molecule can therefore be represented by a state vector $\vec{x} = (x_1, x_2, \ldots, x_{\mathbf{n}})$ where each titratable site i is specified by its protonation state $x_i = 0, 1$. The total number of protonation microstates is therefore $2^{\mathbf{n}}$. The probability for a given site to be protonated is related to the appropriate Boltzmann average over all possible microstates. Because proteins may have tens or even hundreds of interacting sites, the straightforward summation

*Correspondence to: Alexey Onufriev, Department of Computer Science, 660 McBryde, Virginia Tech, Blacksburg, VA 24061. E-mail: onufriev@cs.vt.edu

Received 2 June 2005; Revised 9 November 2005; Accepted 30 November 2005

Published online 21 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20922

involved in the estimation of the partition function becomes a serious problem, in general insurmountable by performing the brute-force addition over all possible states. Without additional approximations, the computational complexity of the problem is exponential, the computation time $\sim 2^{\mathbf{n}}$**.** To solve the problem, various approaches were proposed. In one of the first techniques, the reduced site approximation (REDTI) of Bashford and colleagues,[33] the sites that do not change their protonation significantly at a given pH are regarded as fixed, and the partition function sum is taken over the reduced set of sites whose protonation does change considerably at this pH. The method is highly accurate, but is still exponentially complex,[33] with computation time $\sim 2^{\mathbf{n}}$**.** As a result, realistic computations with this method are limited to molecules with no more than $\sim$ 30 sites. A practical solution to the problem was proposed by Beroza and colleagues,[20] who developed a Monte Carlo estimation whose computational cost scales with the number of sites as $\mathbf{n}^2$. A different type of approach was proposed by Gilson[35]: the groups are separated to form clusters of strongly interacting sites, and their partition functions are calculated exactly, while the interaction between clusters is treated approximately, in a mean-field way. The procedure is initiated by setting each group in its fully ionized state, and then proceeds iteratively to self-consistency. The method was reported to be highly accurate, but the iterations slowed down significantly as the cluster size grew beyond $\sim$ 10 groups.[35] A similar approach was also developed by Yang and colleagues.[22] Approaches that used subsets of titratable groups were also proposed by Lee and coworkers[36] and later by Nielsen.[37] Recently, Spassov and colleagues[27] proposed a very general clustering approach (iterative mobile clustering) to treat additional complexity introduced by conformational flexibility.

In this work, we have asked the following question: can a very simple algorithm based on general clustering ideas lead to estimates of p*K* values of protein ionizable sites well within the generally accepted accuracy of the continuum solvent based approaches? If yes, such an algorithm is likely to be fast and robust, and may therefore be very useful in situations when a large number of structures need to be processed quickly, as in bioinformatics applications, or if an estimation needs to be done in "real time", for example, via a web server. Two such algorithms are considered here. In both cases, the clusters are formed according to the strength of electrostatic site–site interactions: a cluster includes all sites whose interaction with at least one other site is above a predefined threshold. All interactions below the threshold are set to zero. In the first method, *global clustering,* all pairs of sites are considered on an equal footing. Depending on the particular value of the threshold and details of site–site interactions, there may be one or many disjoint, noninteracting clusters in the molecule. The partition function is therefore reduced to a product of sums over a smaller number of states. Alternatively, one focuses on a particular site and considers only the interactions that involve this site, an approach termed *local clustering.* A single cluster is formed in this case, but the procedure must be repeated independently for each titratable site in the molecule. In either case, the computational complexity of the problem is reduced to $\sim \mathbf{n}2^{\mathbf{n}_{\max}}$, where $\mathbf{n}_{\max}$ is the maximum cluster size allowed.

This article begins with a brief review of the continuum electrostatic methods used for calculating the energies of the protonation states. The computation steps of the two clustering methods, global and local, are then formalized. An evaluation of the accuracy and computational efficiency of the methods is presented. The details of the methodology and its use in the implementation and testing of the clustering methods are also described.

THEORY

The association reaction for a protonation microstate $\vec{x}$ is defined as

$$A + H(\vec{x}) \rightarrow AH(\vec{x}) \quad (1)$$

where A is the fully deprotonated form of the macromolecule, $H(\vec{x})$ represents the protons in $\vec{x}$, and $AH(\vec{x})$ is the macromolecule in protonation state $\vec{x}$.[33] The free energy of the given state $\vec{x}$ is evaluated using the following equation:

$$\Delta G(\vec{x}) = \sum_{i=1}^{n} x_i 2.303(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},i}) + \frac{1}{2}\sum_{i,j=1}^{n} W_{i,j}(q_i^0 + x_i)(q_j^0 + x_j) \quad (2)$$

Here, $\mathrm{p}K_{\mathrm{int},i}$ is the intrinsic $\mathrm{p}K_a$ value of the ionizable site i, that is, the pK value that site i would have if all other titratable sites were in the reference state (usually the uncharged state); $W_{i,j}$ is the matrix of the electrostatic interactions between unit charges placed at sites i and j. Also, q_i^0 and q_j^0 define the reference charge values of the associated sites. The equilibrium fractional protonation of site i is

$$\theta_i = \frac{\sum_{\{\vec{x}\}} x_i e^{-\beta \Delta G(\vec{x})}}{\sum_{\{\vec{x}\}} e^{-\beta \Delta G(\vec{x})}} \quad (3)$$

where $\beta = 1/kT$, and $\{\vec{x}\}$ indicates summation over all states of the protonation vector $\vec{x}$. The values θ_i (pH) define the titration curve of site i, which is often characterized by its $\mathrm{p}K_{1/2}$, which is the pH value at which $\theta_i = 1/2$. For an arbitrary set of non-zero W_{ij}, the sums in Equation 3 must be performed over all $2^{\mathbf{n}}$ possible protonation states, hence the computational complexity of the problem discussed above. However, suppose that for a given i, a certain subset of W_{ij} can be assumed to be negligible, $W_{ij} \approx 0$. Then, the corresponding terms in both the numerator and denominator factor out and cancel, leading to a summation over a reduced number of states. Physically this means that sites which do not interact with the site in question do not affect its protonation state and can be ignored. Below we present the details of the two methods to form clusters of such noninteracting sites.

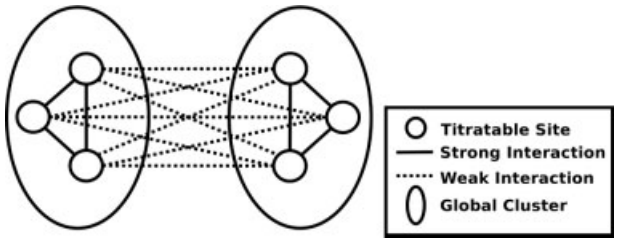


Figure 1. Global clustering. Definition. Each cluster contains only pairs of sites with pair interaction strength (solid short lines) above a predefined threshold. Sites with weak interactions (long dashed lines) between them belong to different clusters.

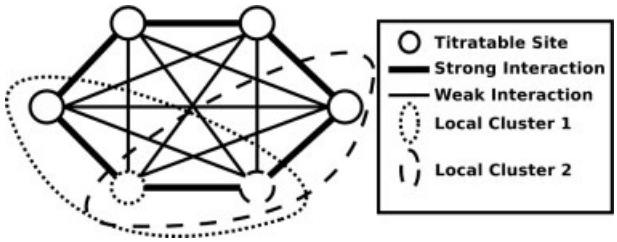


Figure 2. Local clustering. Definition. The dashed and dotted titratable sites are two sites around which the two distinct local clusters are formed. All pairwise interactions above the preset threshold value are shown by thick short lines, thin lines represent interactions below the threshold. Each site whose (direct, pairwise) interaction strength with the chosen site is above the threshold belongs to the given local cluster. In this example the size of the local cluster is 3, and only the nearest neighbor titratable sites form clusters. Note that local clusters may overlap.

Global Clustering

Clusters are defined as subsets of titratable sites based on pairwise interaction strength between members greater than a threshold value. With a threshold value W_{th} defined, a cluster is a subset of titratable sites such that there exists a path between every pair of titratable sites in the cluster where every interaction strength along the path is greater than the threshold. If there does not exist a path of interactions between two sites such that every interaction strength is greater than the threshold, the two sites do not exist in the same cluster. Sites belonging to different clusters are considered as noninteracting (Fig. 1). The calculation of protonation for a site i in a cluster γ is done via:

$$\theta_{i\in\gamma} = \frac{\Sigma_{\{\vec{x}_\gamma\}} x_i \exp\left[\Sigma_{j\in\gamma} x_j 2.303(\mathrm{pH} - \mathrm{p}K_{\mathrm{int}_j}) + \frac{1}{2}\Sigma_{\alpha,\beta\in\gamma} W_{\alpha,\beta}(q^0_\alpha + x_\alpha)(q^0_\beta + x_\beta)\right]}{\Sigma_{\{\vec{x}_\gamma\}} \exp\left[\Sigma_{j\in\gamma} x_j 2.303(\mathrm{pH} - \mathrm{p}K_{\mathrm{int}_j}) + \frac{1}{2}\Sigma_{\alpha,\beta\in\gamma} W_{\alpha,\beta}(q^0_\alpha + x_\alpha)(q^0_\beta + x_\beta)\right]} \quad (4)$$

where $\{\vec{x}_\gamma\}$ is the set of all possible vectors over the cluster of sites γ and W is a matrix where $W_{\alpha,\beta}$ is the interaction strength between titratable sites α and β. The size of each cluster controls both the accuracy of the approximation and the computational speed, see below. After the global clusters are defined, the free energy and titration curves are calculated for the set of sites in each cluster independently.

Because it is the maximum cluster size $\mathbf{n}_{\max}$ that directly controls the computational speed, we use this parameter, rather than the interaction energy threshold, to define global clusters. In practice, the threshold is initially set infinitely small to allow for all titratable sites to exist in a single cluster. The threshold is then incremented to separate the titratable sites into distinct clusters; the procedure stops when the maximum cluster size has reached the desired threshold. A brute-force computation of the titration curve via Equation 3 can then be applied to the individual clusters, the calculation time will be exponentially proportional to each cluster size, which is less than $\mathbf{n}_{\max}$ by construction. In practice, we apply the reduced site approximation (REDTI) to evaluate Equation 3; this approximation effectively decouples sites that titrate at very different pK, thus reducing the computational time even further.

Figure 1 illustrates the idea behind the global clustering. In this example, there are six titratable sites which yields $2^6 = 64$ possible protonation microstates. The lines between each site represent the electrostatic interaction strength between the sites. Here, two clusters are formed where the electrostatic interactions are strong between all the sites within each cluster and weak between all the sites in different clusters. In this example, a partition function computation on the entire set of sites would require 64 evaluations. The global cluster approximation requires two sets of evaluations each of size 2^3, which sums to only 16 total evaluations. With the maximum cluster size $\mathbf{n}_{\max}$ fixed, the time savings become dramatic for systems with a large number of sites.

Local Clustering

The second method of clustering considered here, referred to as local clustering, defines each individual cluster around the site of interest, and is used only to evaluate the titration curves of this site. To estimate protonation states of all sites in the molecule, the procedure must be repeated for each site. Again, only those sites whose pairwise interactions with the given site are above the threshold are kept in the cluster, and all other interactions are ignored. The procedure used to construct the cluster is similar to that for the global clustering described above, except there is only one cluster per site and clusters centered around different sites may overlap (see an example in Fig. 2). The calculation of protonation for site i is based exclusively on the sites defined in the cluster associated with site i, cluster γ_i.

$$\theta_i = \frac{\Sigma_{\{\bar{x}_{\gamma_i}\}} x_i \exp\left[\Sigma_{j\in\gamma_i} x_j 2.303(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},j}) + \frac{1}{2}\Sigma_{\alpha,\beta\in\gamma_i} W_{\alpha,\beta}(q_\alpha^0 + x_\alpha)(q_\beta^0 + x_\beta)\right]}{\Sigma_{\{x_{\gamma_i}\}} \exp\left[\Sigma_{j\in\gamma_i} x_j 2.303(\mathrm{pH} - \mathrm{p}K_{\mathrm{int},j}) + \frac{1}{2}\Sigma_{\alpha,\beta\in\gamma_i} W_{\alpha,\beta}(q_\alpha^0 + x_\alpha)(q_\beta^0 + x_\beta)\right]} \quad (5)$$

Later we will see that a very reasonable accuracy can be achieved with a maximum cluster size set less or equal to 20, making the associated computational costs negligible compared to the typical costs involved in estimating the charge–charge interactions. In practice, we use the clustering method in conjunction with REDTI to further reduce computational time.

RESULTS AND DISCUSSION

To estimate the accuracy of the clustering methods discussed above, we compute p$K_{1/2}$ values for every titratable site in a test set of 602 proteins, representing a wide variety of protein sizes and folds.[34] The total number of titratable sites analyzed is ~ 20,000. The resulting p$K_{1/2}$ values are compared to the reference values obtained by the established Monte Carlo methodology[20] developed by P. Beroza and colleagues, conveniently available as the software package MCTI. Note that a direct comparison with the exact partition sum, Equation 3, is not only impractical, but computationally unrealistic for molecules with more than ~ 35 titratable sites. In what follows we assume that the site–site interaction matrix, $W_{i,j}$, and the pK_{intr} values that enter Equation 2 for each test structure have been precalculated. We use the continuum electrostatics methodology (see Computational Methods section) for these computations, although the algorithms proposed here can be used with any other method of estimating the terms in Equation 2.

Details of how the maximum cluster size $\mathbf{n}_{\max}$ affects the average accuracy of the p$K_{1/2}$ computations are presented in Figure 3. Both methods show curves of increasing accuracy as the maximum cluster size is increased. The local method is the more accurate of the two: if the five largest proteins (with more than 120 titratable sites) out of 602 tested are excluded, the average error of the method is less than 0.15 p*K* units for a maximum cluster size of 20. If all tested proteins (a total of ~ 20,000 sites) are taken into account, the average error is still only about 0.25 p*K* units. A more detailed analysis, Figure 4, reveals that within the 95% confidence interval, the error introduced by the local method at $\mathbf{n}_{\max} \sim 20$ is guaranteed to be no larger than 0.25 if five largest proteins are excluded. If all 602 proteins are included, this error is no larger than 0.3. This number is a statistical upper bound and is much larger than the errors one can expect for a "typical" medium-size protein: Table I shows estimated p$K_{1/2}$ for a set of ionizable residues in lysozyme (HEWL) and BPTI. These proteins were used many times before in the analysis of the accuracy of p$K_{1/2}$ predicting methodology.[38] Some titratable groups exhibit considerable shifts in p$K_{1/2}$ relative to the model compound values, and were proposed[39] to serve as discriminative benchmarks for testing of the p*K*-prediction methods. Note the maximum difference of only 0.032 for BPTI and 0.083 for HEWL in p$K_{1/2}$ predicted by the local clustering (maximum cluster size = 17) methods relative to the Monte Carlo standard. In fact, for all 261 proteins in the test group with 20 to 39 titratable sites per protein, the maximum error is 0.10, averaged over the entire group. The advantage offered by the local clustering method is a two orders of magnitude shorter run time. Note that the error introduced by the (local) clustering approach is considerably smaller than the overall deviation of the predicted p$K_{1/2}$ s from experiment. While the predicted values still compare reasonably well to experiment, this agreement (or disagreement) characterizes the method of estimating the free energy terms in the fundamental Equation 2 rather than the algorithms for computing the protonation partition function discussed here. For illustration purposes, here we have used the relatively simple single-conformation continuum solvent methodology to estimate the relative free energies of protonation microstates. The use of more sophisticated methods, example multiconformer models[30] or semimicroscopic approaches[25] that go beyond the continuum approximation, is likely to improve the accuracy of the computed free energies and hence the agreement with the experiment. The clustering algorithms developed here may still be useful in this case as the postprocessing step if a fast evaluation of the partition function is desirable. The accuracy of the local clustering approach may, perhaps, be unexpected, given that the approximation neglects a large number of site–site interactions. This can be explained by the fact that the electrostatic interactions removed from consideration by the local clustering algorithm have relatively small values. For example, when the local clustering algorithm with $\mathbf{n}_{\max} = 20$ is applied to the largest group of proteins in the test set—those with the number of titratable sites in the range 20 to 39—the maximum interaction W_{ij} ignored is only 0.014 kcal/mol $\ll$ kT $\simeq$ 0.6 kcal/mol. For the largest proteins tested, those in the group with 80 to 179 titratable sites, the strongest interactions neglected are still less than 0.21 kcal/mol. However, the accuracy of the computed p*K*s is becoming worse as larger interaction are being neglected; the method may become impractical for the few largest proteins tested, Figure 4. See Methods for details of how these values have been computed. The global clustering method is reasonably accurate for proteins with the number of titratable sites fewer than 40 (test subgroups 1 and 2): the upper error bound is ~ 0.3 in this case, Figure 4. The result is in a qualitative agreement with the conclusions of Nielsen and colleagues,[37] who earlier proposed and tested a method which appears

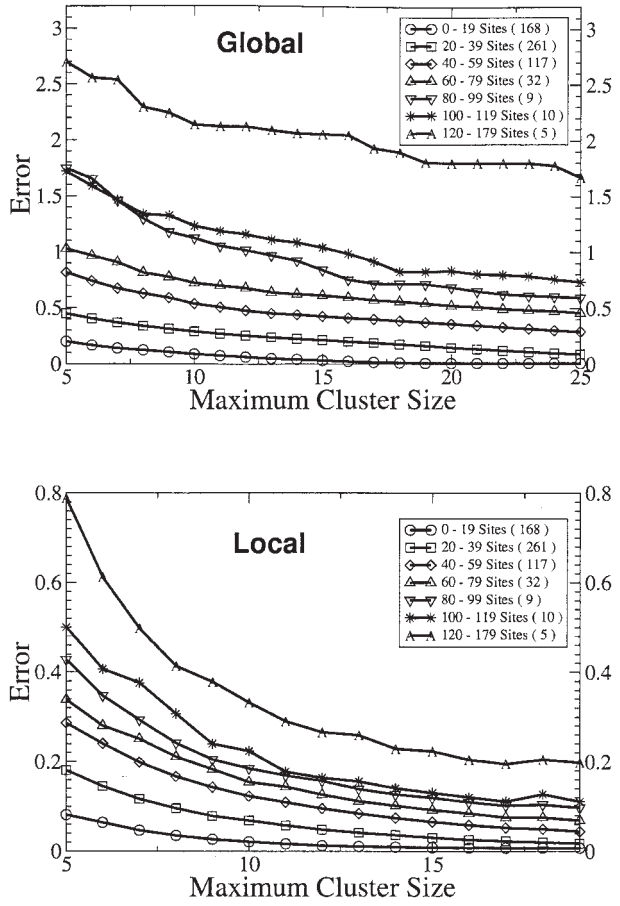


Figure 3. The average absolute error in p$K_{1/2}$ values produced by the two clustering methods as a function of the maximum cluster size (top, global clustering; bottom, local clustering). The errors are computed relative to the corresponding p$K_{1/2}$ values obtained by the Monte Carlo method used as a reference. The test set of 602 proteins is broken into subsets based on the number of titratable sites found in each protein; these are shown in the legend box, along with the number of proteins in each subset (in parenthesis).

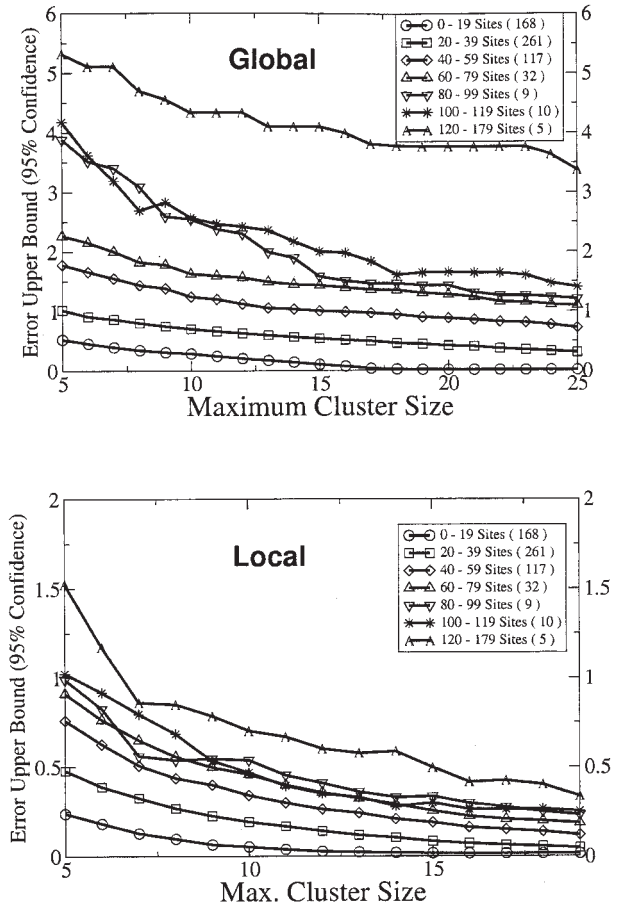


Figure 4. The 95% confidence error range for the global method (top) and local method (bottom) for each protein as a function of the maximum cluster size. The errors are computed relative to the corresponding p$K_{1/2}$ values obtained by the Monte Carlo method. The proteins are grouped in subsets by the number of titratable sites as in Figure 3.

similar to the global clustering approach of the current work. The test cases used by Nielsen and colleagues were lysozyme (33 sites) and xylanase (40 sites). Our analysis shows, Figure 4, that the global clustering algorithm does not perform as well for larger proteins: the upper bound of the error can be larger than 1 pK unit for many proteins, and even exceed 4 units for some (subgroup with number of sites $>$ 120.). We therefore do not recommend it as a substitute for the local clustering or Monte Carlo approach when accurate estimates are required for proteins with number of titratable sites $>$ 40; however we will see later that the global method may be useful in a very different context.

The execution time of the clustering algorithms for any protein depends on two factors: the number of titratable sites and the maximum cluster size allowed, Figures 5 and 6. The relative execution times of the two clustering methods and the Monte Carlo method in terms of the number of titratable sites are presented in Figure 6. The execution time is the total run time it takes to obtain the p$K_{1/2}$ values for each titratable site in a given protein, starting from the precalculated values of the site–site interaction matrix W and intrinsic pK_{intr} values. The time includes the (almost negligible) time of forming the cluster(s) and the execution time of running REDTI on every cluster.

As seen in Figure 6, the time it takes to compute a set of p$K_{1/2}$ values using the clustering methods increases significantly with the number of sites in the protein, but both the global and the local clustering methods run substantially faster than the reference Monte Carlo method (with typical settings) for the same number of sites. The relative speed-up is significant, almost 2 orders of magnitude for proteins with less than 70 titratable sites, and at least 1 order of magnitude for the very large proteins with a number of sites greater than 70. Note that 90% of the proteins tested have fewer than 70 titratable groups, and because the test set we have used is rather large and representative, in most cases one can expect a speed-up of about 100 (relative to Monte Carlo method) for a randomly selected protein.

Of course, computation of the partition function is just the last step in a typical p$K_{1/2}$ predicting protocol: comput-

TABLE I. An Example of the Local Clustering Algorithm Performance Detailed for Two Typical Proteins (BPTI and hen Egg-White Lysozyme) Often Used for the Purpose of Testing p*K* Predicting Methods

BPTI	Speed Up: 69			
Site	Local Cluster p$K_{1/2}$	Monte Carlo p$K_{1/2}$	Difference	Experimental pK
Nter1	6.301	6.304	0.003	8.1
Asp3	3.453	3.460	0.007	3.4
Glu7	5.671	5.671	0.000	3.7
Lys15	10.385	10.384	0.001	10.6
Lys26	10.423	10.423	0.000	10.6
Lys41	10.182	10.214	0.032	10.8
Lys46	9.768	9.772	0.004	10.6
Glu49	4.158	4.156	0.002	3.8
Asp50	2.404	2.402	0.002	3.0
Cter58	3.788	3.792	0.004	2.9
HEWL	Speed Up: 110			
Site	Local Cluster p$K_{1/2}$	Monte Carlo p$K_{1/2}$	Difference	Experimental pK
Nter1	6.771	6.742	0.029	7.90
Lys1	9.646	9.630	0.016	10.80
Glu7	3.344	3.330	0.014	2.85
Lys13	10.438	10.434	0.004	10.50
His15	5.605	5.582	0.023	5.36
Asp18	1.652	1.602	0.050	2.66
Tyr20	15.723	15.760	0.037	10.30
Tyr23	10.673	10.650	0.023	9.80
Lys33	11.000	10.991	0.009	10.60
Glu35	5.119	5.061	0.058	6.20
Asp48	−0.629	−0.660	0.031	1.60
Asp52	2.095	2.012	0.083	3.68
Tyr53	26.156	25.184	0.028	12.10
Asp66	−2.218	−2.247	0.029	0.90
Asp87	0.967	0.926	0.041	2.07
Lys96	11.092	11.088	0.004	10.80
Lys97	10.816	10.789	0.027	10.30
Asp101	4.603	4.582	0.021	4.09
Lys116	9.154	9.159	0.005	10.40
Asp119	3.318	3.293	0.025	3.20
Cter129	3.640	3.631	0.009	2.75

The p$K_{1/2}$ values of the ionizable groups are computed by the proposed local clustering algorithm ($\mathbf{n}_{max} = 17$) and compared to those calculated by the conventional Monte Carlo based (MCTI) method: the two methods produce nearly identical values (the difference is shown in the fourth column), while the clustering algorithm is 60 to 100 times faster. The speed-up is estimated as ratio of the corresponding execution times. The pK_{intr} and W_{ij} values required to compute p$K_{int,i}$ via Equations 2 and 3 have been precomputed using the standard continuum solvent based methodology (MEAD), see Methods. For reference, corresponding experimental pKs are listed where available.

ing the terms in Equation 2 may also be time-consuming, depending on the particular methodology employed. Still, the use of a clustering method may offer a significant advantage. For example, for the largest subgroup of our test proteins (20–39 sites), the average execution time of the Monte Carlo step is 93 s, while finite-difference PB calculations of W_{ij} and pK_{intr} by MEAD package take 76 s on average. Reducing the time of the first step by an order of magnitude or more will half the total run time. The speed-up is even larger if the finite difference routine is replaced by one of the faster implicit solvent methods such as the generalized Born.[40] the time it takes to estimate W_{ij} and pK_{intr} is then reduced to 24 s, and the overall speed-up is about a factor of 5 on average for the above group of proteins.

The practical usefulness of a computational method is largely determined by two factors: its accuracy and the associated computational costs. For the clustering approach considered here, the computational time for any protein is directly related to the maximum cluster size allowed, Figure 5. For the local method, the dependence is clearly monotonic and seemingly exponential, as expected. The execution time also grows fast for the global clustering, although the dependence on the maximum cluster size is not always monotonic, Figure 5(b). This is likely because the global algorithm does not necessarily produce clusters of maximum allowed size; depending on the distribution of sites and the interaction strengths, the largest cluster may happen to be smaller than the maximum allowed.

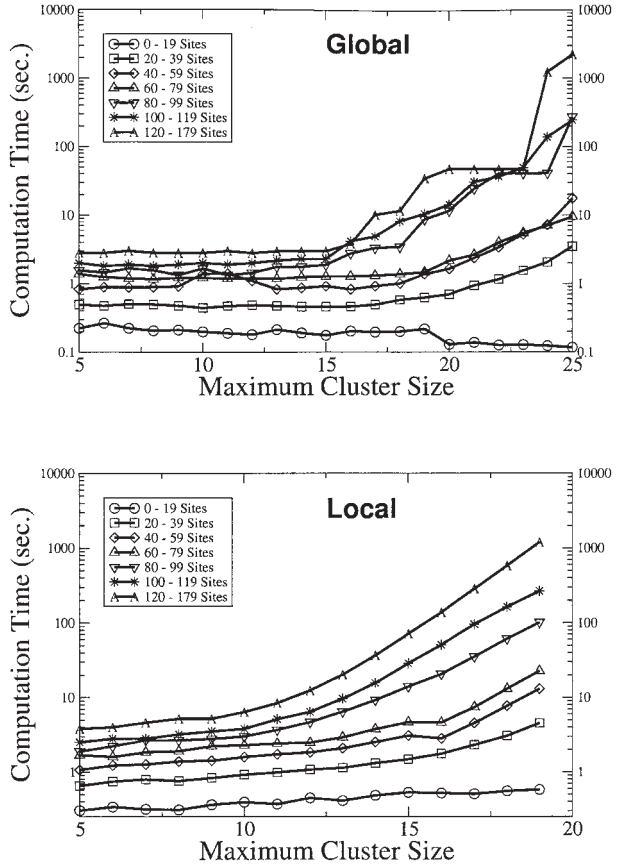


Figure 5. Computation times of the two clustering methods as a function of the maximum cluster size. The computational time of the local clustering algorithm (bottom) grows almost monotonically (exponentially) with the maximum cluster size, while larger variations in this dependence are observed with the global clustering method (top). The computation time is defined as the total run time it takes to obtain a set of all $pK_{1/2}$ values for a given protein, starting from the precalculated values of the site–site interaction matrix W_{ij} and intrinsic pK_{intr} values.

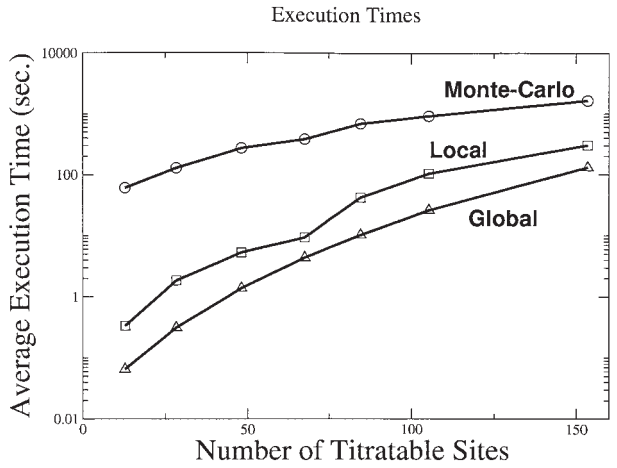


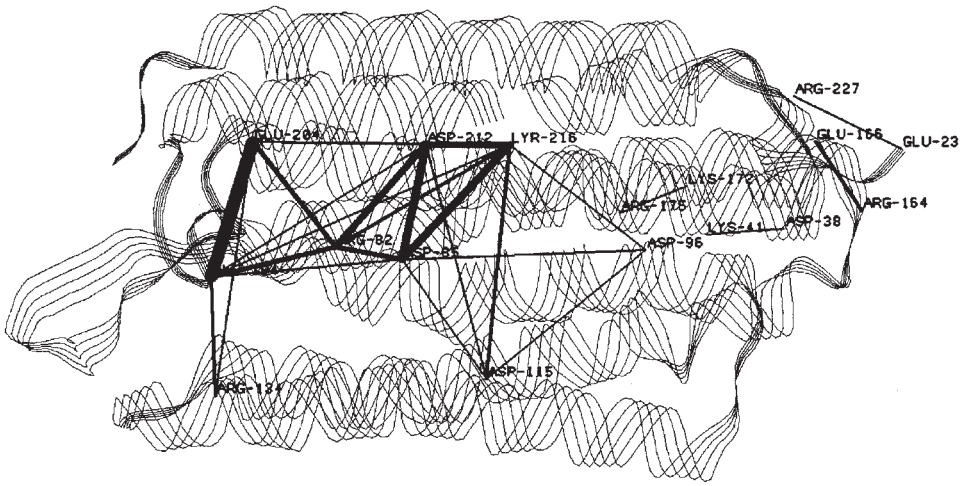
Figure 6. Average execution time of the calculations based on the two clustering algorithms and the Monte Carlo method (MCTI) for a set of 602 proteins. The execution time is the total run time it takes to obtain a set of all $pK_{1/2}$ values for a given protein, starting from the precalculated values of the site–site interaction matrix W_{ij} and intrinsic pK_{intr} values. The maximum cluster size is 17 in all cases. Parameters of the MCTI program are set to their default values; see Methods.

So far, it has not been clear which of the two clustering methods is more useful, from a practical standpoint, in calculating protein titrating curves, especially for smaller proteins. To decide, we compare Figures 5 and 4. The execution of the local method with maximum cluster size of 17 produces results which are both more accurate and require less execution time than the execution of the global method with maximum cluster size of 25. Therefore, if the goal is accurate calculation of the titration curves (or $pK_{1/2}$ values), the local clustering method appears to be more practical. The only exception may be small proteins, for which the error of the global method, although still larger than that of the local one, is still quite small compared to other errors involved. At the same item, the global clustering method may become useful for addressing other important questions, as illustrated in the next section.

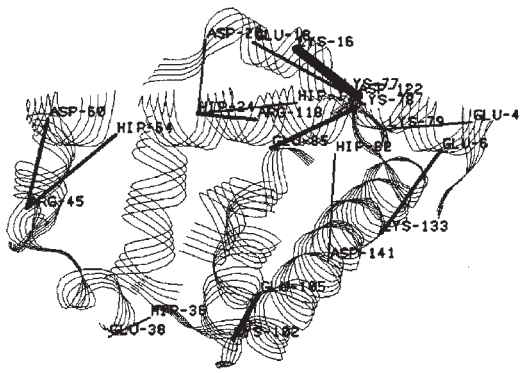
Clusters of Strongly Coupled Sites and Biological Function

Understanding the precise relationships between the structure of a biomolecule and its function is one of the key challenges in the biocomputational field. Specific structural signatures, such as mutual positions of key residues in active sites, were used to predict molecular function.[41,42] Unusual pK_a values and titration curves of ionizable groups in active sites were correlated with specific enzymatic activity.[16] Here, we put forward a conjecture that the presence of a (global) cluster of strongly coupled ionizable groups within a protein may be indicative of the cluster's importance to the function of the protein. A proof of this statement would require a detailed analysis of a large, statistically significant test-set, and is beyond the scope of this purely methodological work. We therefore limit ourselves to a relevant example that supports the above conjecture; the main goal is to present the tools that can be used for this kind of analysis.

The light-driven proton pump, *bacteriorhodopsin,* is a relatively small protein made up of seven membrane-spanning helices and a retinal chromophore bound to Lys216 by a Schiff-base linkage in the central part of the molecule.[12] Absorption of a light quantum by the chromophore triggers a series of retinal isomerization changes, protonation/deprotonation events and protein structural changes that comprise the bacteriorhodopsin resulting in the net transfer of one proton from the cytoplasmic side to the extracellular side of the membrane. A global clustering analysis for this protein is presented in Figure 7(a), where we have used the interaction coupling strength of W_{th} = 2.21 kcal/mol ($\sim$ 3 kT) as the threshold. The analysis reveals one large global cluster of eight sites (and a few isolated clusters, two sites in each). The large global cluster contains the following sites: Lys-216/Retinal Schiff Base (Lyr-216), Asp96, Asp115, Asp85, Asp212, Arg82, Glu194, and Glu204. This happens to be the key, highly conserved group of residues determining the proton-pumping function of bacteriorhodopsin. It is also known that the function of the protein is quite robust to mutations



(a) Bacteriorhodopsin



(b) Myoglobin

Figure 7. Global clusters in bacteriorhodopsin (top) and myoglobin (bottom). The thick lines show pairwise electrostatic interactions whose strength is larger than 2.12 kcal/mol; the width of each line represents the relative strength of each interaction. In bacteriorhodopsin this threshold value defines a single large global cluster of nine sites, all of which were implicated earlier as important for the vectorial proton transport, which is the function of bacteriorhodopsin. The strong interaction between these site was found[12] critical for the protein function. In contrast, no large cluster of strongly coupled sites is found in myoglobin when the same threshold value, 2.12 kcal/mol is used to produce the diagram shown at the bottom. The biological function of myoglobin—oxygen transport and storage—does not depend on the presence of a cluster of strongly coupled groups. The ribbon diagrams are produced by RasMol.

of amino-acids outside of this cluster.[43–45] Doubling the threshold value leaves the cluster almost unaffected (ASP-96 disappears), while decreasing it by a factor of two does not bring about new clusters of more than two sites in each. The strong electrostatic coupling between the above key groups was shown to be crucial for the function of the protein;[12] the proton transfer mechanism requires that a change in the protonation state of one residue in the cluster must have a significant effect on the proton affinity of others along the proton transport chain.

A very different scenario with respect to global clustering is found in an oxygen transport protein myoglobin, Figure 7. At the same value of the site–site coupling threshold, 2.12 kcal/mol, no large clusters are present. We note that the biological function of myoglobin does not require the presence of a strongly coupled chain of titratable groups in the protein.

CONCLUSION

In this article we have addressed the problem of time-effective computation of partition sums over an exponentially large number of protonation microstates; the problem arises in most theoretical algorithms that use principles of statistical thermodynamics to predict p*K*s of ionizable groups based on molecular structure. As the number of protonation microstates depend exponentially on the num-

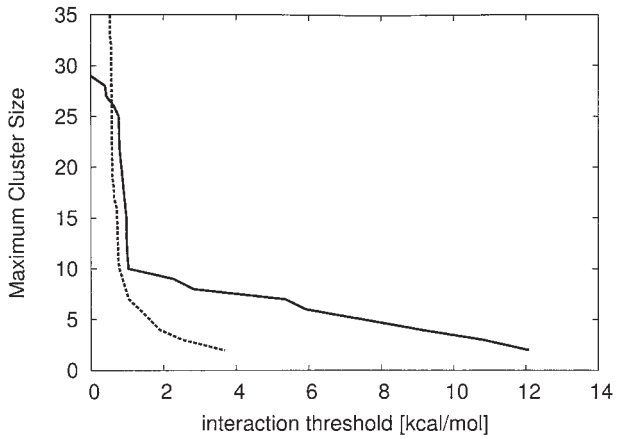


Figure 8. The maximum cluster size as a function of the interaction threshold for *bacteriorhodopsin* (solid line) and *myoglobin* (dashed line). The global clustering algorithm is used. The long tail of the curve for bacteriorhodopsin indicates the presence of a large cluster of strongly coupled ionizable groups.

ber of ionizable groups present in the molecule, direct computation of the sum is not realistically feasible for many typical proteins that may have tens or even hundreds of ionizable groups.

We have tested a straightforward and computationally robust approximate algorithm for computing these sums for proteins. The algorithm subdivides the interacting sites into independent clusters, based on the strengths of precomputed site–site interactions: interactions above a preset threshold are kept, and all others are completely neglected. The resulting partition function is factorizable into computationally manageable components. Two variants of the approach are discussed. The first approach forms global clusters in which all titratable sites are treated on an equal footing; the other variant forms clusters centered around a chosen site (local clustering), and the procedure is repeated for each site. A single parameter—the maximum allowed cluster size—controls both the accuracy and the computational time of the procedure. The algorithms are validated on a representative test set of 602 proteins with the combined total of $\sim$ 20,000 ionizable sites. The the p$K_{1/2}$ values computed by each of the proposed methods are compared with those obtained by the standard Monte Carlo approach traditionally used in this context. The main, perhaps somewhat surprising, conclusion is that choosing a fairly modest maximum allowed cluster size, $\sim$ 20, is enough to produce results (local clustering) accurate to within about 0.25 pK units relative to the Monte Carlo reference for the vast majority of proteins tested. The corresponding execution time is between one to two orders of magnitude smaller than that of the standard Monte Carlo method: for the vast majority of the tested proteins it is less than 10 s on a typical desktop computer. Given that the consensus accuracy of pK estimates by the continuum solvent methodology is $\sim$ 1pK unit, it makes the proposed local clustering approximation a reasonable, practical approach to be used as the end-step of pK predictive methods. The robustness and speed of the approach stem from its algorithmic simplicity which in turn is the consequence of complete neglect of all the site–site interactions bellow a preset threshold value. While other, more accurate/complex approaches based on the clustering idea were explored before, our finding that the very minimal version of it is accurate enough for practical purposes may be of value to the computational community. In fact, the local clustering algorithm described here is already being used in an automated web-based server[46] (http://biophysics.cs.vt.edu/H++) that predicts the pK of ionizable groups in biomolecules. The speed and robustness of the approach proved especially beneficial in the context of real-time web-based computations. While the pK calculations reported here are based on a particularly straightforward single-conformer continuum electrostatics approach, the algorithms used to compute the protonation partition function are general, and can be use with other, more sophisticated methods of estimating pKs in macromolecules.

Another observation we have made is that the distribution of the global clusters in a biomolecule might correlate with its biological function. The two examples considered here, *bacteriorhodopsin* and *myoglobin* show qualitatively different behavior in this respect. In bacteriorhodopsin, a large cluster of electrostatically strongly coupled groups is found, and all of these groups are known to be important for the protein's main function: light-induced proton transport. In contrast, the main biological function of myoglobin—oxygen storage and transport—does not require strongly coupled clusters, and these are indeed not found in the protein. The ability to explore electrostatic site–site coupling visually, in the context of macromolecular structure, may facilitate understanding of the structure–function relationships. Clusters of electrostatically coupled groups, along with the relative strengths of the corresponding interactions, can be easily visualized using standard structure-viewing packages: we have prepared a set of programs necessary for such visualization, they are free to download from http://www.cs.vt.edu/~onufriev/software.html. While the global clustering algorithm is well suited for identification of clusters of strongly coupled sites, we do not recommend to use it for accurate computation of pK in proteins with more than 40 titratable groups.

COMPUTATIONAL METHODS

Structures

We have used a set of 602 representative proteins[34] for validation of the clustering algorithms presented. Missing heavy atoms and protons (assuming standard protonation states of titratable groups) are added, and atomic partial charges and radii (Bondi) are assigned using the protonate and LEAP modules of AMBER-8. The structure of *bacteriorhodopsin* (PDB ID 1QHJ) has been prepared as described in Reference 12, and includes a model of lipid bilayer.

Electrostatic Calculations

The continuum electrostatics methodology widely used to calculate the energetics of proton transfer is described elsewhere;[9,47] the model is available in the free software package MEAD.[48] The protein is treated as a low dielectric

medium $\epsilon_{in} = 4$, while the surrounding solvent is assigned a high dielectric constant $\epsilon_{out} = 80$. The electrostatic screening effects of (monovalent) salt enter via the Debye–Huckel screening parameter $\kappa = 0.1\text{Å}^{-1}$, which roughly corresponds to physiological conditions.

The difference between a sidechain's pK_a and the pK_a of the corresponding model compound in free solution is determined by the combined effect of two distinct contributions to the total electrostatic (free) energy change. First is the Born term or desolvation penalty, which always penalizes burial of a charge inside a low dielectric medium. Second is the background term which represents the electrostatic interactions of the group in question with all other fixed charges in the molecule not belonging to any titratable groups. These energy terms, as well as the matrix of site–site interactions W_{ij} are estimated through a sequence of finite-difference Poisson–Boltzmann calculations in which sites in the protein and their corresponding model compounds have their charge distributions set to those of the protonated or deprotonated form, and suitable energy differences are taken. The computations are performed by the multiflex module of MEAD (single-conformer regime). In the finite difference lattices, two levels of focusing are used. In the coarsest level the bounding box is set to twice the molecule's maximum extent and the grid points are spaced 2 Å apart. The finest lattice is "focused" on the region of interest, and the grid points are 0.5 Å apart. The probe radius for defining the molecular surface, which is used as the boundary between the interior and exterior dielectric regions, is set to 1.4 Å. For the protonated states of ASP and GLU, in which the correct location of the proton is not known a priori a "smeared charge" representation is employed, in which the neutralizing positive charge is symmetrically distributed: 0.45 on each carbonyl oxygen atom, and 0.1 on the carbon atom. The maximum values of W_{ij} ignored by the local clustering approach have been computed as follows: for each protein in the test group, max $\{W_{ij}\}$ neglected for a given $\mathbf{n}_{\max}$ is computed for all of its titratable sites; the number is then averaged over all of the proteins in the group.

Calculation of Titration Curves and p$K_{1/2}$ Values

The electrostatic calculations outlined above provide (free) energies of each of the $2^{\mathbf{n}}$ protonation microstates[12] in the system where $\mathbf{n}$ is the number of ionizable sites. As mentioned above, direct computation of the the partition sums (and p$K_{1/2}$) is not feasible for $\mathbf{n}$ larger than about 30. We have therefore relied on the established Monte Carlo procedure to produce a set of p$K_{1/2}$ values to be used as reference when validating the clustering approximation described here. The Monte Carlo approach is implemented as the freely available program MCTI due to P. Beroza.[49] Each MC run is performed with 1000 full steps, 10,000 reduced steps, an all-inclusive pH range of -55 to 55, and a tolerance of 0.000001. These are the default settings for the code, they were used earlier in the context of pK calculations.[12]

Sites whose computed p$K_{1/2}$ values outside of the $\pm$ 5 interval around the pH value at which the experimental structure was obtained (this information is retrieved from the PDB record) are excluded from the analysis. The rationale for the exclusion is that the protein structure outside of this range would most likely be very different from the original structure used in the calculations, and the computed p$K_{1/2}$ values are therefore unrealistic. We also exclude the sites whose computed titration curves are non-monotonic: it was shown earlier[50] that p$K_{1/2}$ values for such curves are either undefined or meaningless. About 200 (out of 20,000 total) sites have been excluded on this basis.

Implementation Details

Global clustering

The input data (output of the MEAD electrostatic calculations) is a set of intrinsic pK values for each titratable site and a matrix W_{ij} of the electrostatic site–site interaction strengths over a set of $\mathbf{n}$ titratable sites. The latter can be viewed as a complete graph where each titratable site is represented by a vertex i and the interaction between sites i and j is an edge with an associated strength W_{ij}, see Figure 1 as an example. We begin by sorting the edges based on edge strength, from large to small. Using a disjoint set structure,[51] we place each vertex into an individual set. We then read through the edges from large to small. For each edge (i,j) encountered, a union of the the two associated sets is formed, the set that contains i and the set that contains j. We keep track of the size of each set during the process of forming the union. If a possible union between two sets would produce a set with size greater than the maximum cluster size we terminate the process. At this point, there exists a set of sets (clusters) where each cluster has size less than or equal to the maximum cluster size. For each such cluster, two files are produced: a file containing the intrinsic pK values for the sites in this set only, and the associated matrix of interaction strength. The resulting files are then passed to REDTI for further processing. The overall time to generate the global clusters is $O[\mathbf{n}^2 \log(\mathbf{n})]$.

Local clustering

The input data structure is the same as in the previous section. For each titratable site i, a list of all the site–site interaction strengths in W_{ij} is formed. The list is sorted down, from large to small, and the maximum cluster size number of top elements form a cluster of titratable sites. The corresponding files containing the intrinsic pK values and the associated matrix of interaction strength are produced and passed to REDTI for further processing. The overall time to generate the local clusters is $O[\mathbf{n}^2 \log(\mathbf{n})]$.

Benchmarks and Timings

All computations are performed on a Dell Dimension 4600 series with a 2.66 GHz Intel Pentium 4 processor, 1024 MB DDR SDRAM at 333 MHz, 80 GB 7200 RPM Ultra ATA hard drive, and Red Hat 9.0 operating system.

Software

Both of the clustering algorithms described here, along with additional programs necessary for visualization of

the global clusters, have been implemented as free software packages available to download from http://www.cs.vt.edu/~onufriev/software.html

REFERENCES

1. Perutz M. Electrostatic effects in proteins. Science 1978;201:1187–1191.
2. Warshel A, Russell ST, Churg AK. Macroscopic models for studies of electrostatic interactions in proteins: limitations and applicability. Proc Natl Acad Sci U S A 1984;81:4785–4789.
3. Warshel A, Russell S. Calculations of electrostatic interactions in biological systems and in solutions. Q Rev Biophys 1984;17:283–422.
4. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
5. Davis ME, McCammon JA. Electrostatics in biomolecular structure and dynamics. Chem Rev 1990;94:7684–7692.
6. Baker NA, McCammon JA. Electrostaic interactions. In: Structural bioinformatics. Wiley: New York; 2002.
7. Warshel A, Åqvist J. Electrostatic energy and macromolecular function. Ann Rev Biophys Biophys Chem 1991;20:267–298.
8. Warshel A. Calculations of enzymatic reactions: Calculations of pK_a, proton transfer reactions, and general acid catalysis reactions in enzymes. Biochemistry 1981;20:3167–3177.
9. Fersht A, Shi J, Knill-Jones J, Lowe D, Wilkinson A, Blow D, Brick P, Carter P, Waye M, Winter G. Hydrogen bonding and biological specificity analysed by protein engineering. Nature 1985;314:235–8.
10. Szabo G, Eisenman G, McLaughlin S, Krasne S. Ionic probes of membrane structures. Membrane structure and its biological applications. Ann N Y Acad Sci 1972;195:273–290.
11. Sheinerman FB, Norel R, Honig B. Electrostatic aspects of protein-protein interactions. Curr Opin Struct Biol 2000;10:153–159.
12. Onufriev A, Smondyrev A, Bashford D. Proton affinity changes during unidirectional proton transport in the bacteriorhodopsin photocycle. J Mol Biol 2003;332:1183–1193.
13. Yang AS, Honig B. Electrostatic effects on protein stability. Curr Opin Struct Biol 1992;2:40–45.
14. Pots A, de Jongh H, Gruppen H, Hessing M, Voragen A. The pH dependence of the structural stability of patatin. J Agric Food Chem 1998;46:2546–2553.
15. Whitten S, Garcia-Moreno B. pH dependence of stability of staphyococcal nuclease: evidence of substantial electrostatic interactions in denatured state. Biochemistry 2000;39:14292–14304.
16. Ondrechen M, Clifton J, Ringe D. THEMATICS: A simple computational predictor of enzyme function from structure. Proc Natl Acad Sci U S A 2001;98(22):12473–12478.
17. Tanford C, Kirkwood J. Theory of protein titration curves. J Am Chem Soc 1957;79:5333–5339.
18. Tanford C, Roxby R. Interpretation of protein titration curves. Biochemistry 1972;11:2192–2198.
19. Bashford D, Karplus M. pK_as of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 1990;29(44):10219–10225.
20. Beroza P, Fredkin DR, Okamura MY, Feher G. Protonation of interacting residues in a protein by Monte Carlo method. Proc Natl Acad Sci U S A 1991;88:5804–5808.
21. Takahashi T, Nakamura H, Walda A. Electrostatic forces in two lysozymes: calculations and measurements of histidine pK_a values. Biopolymers 1992;32:897–909.
22. Yang AS, Gunner MR, Sampogna R, Sharp K, Honig B. On the calculation of pK_as in proteins. Proteins 1993;15:252–265.
23. DelBuono G, Figueirido F, Levy R. Intrinsic pK_as of ionizable residues in proteins: an explicit solvent calculation for lysozyme. Proteins 1994;20:85–97.
24. Demchuk E, Wade RC. Improving the continuum dielectric approach to calculating pK_as of ionizable groups in proteins. J Phys Chem 1996;100:17373–17387.
25. Sham YY, Chu ZT, Warshel A. Consistent calculations of pK_as of ionizable residues in proteins: Semi-microscopic and microscopic approaches. J Phys Chem 1997;101:4458–4472.
26. Ullmann GM, Knapp EW. Electrostatic models for computing protonation and redox equilibria in proteins. Eur Biophys J 1999;28:533–551.
27. Spassov VZ, Bashford D. Multiple-site ligand binding to flexible macromolecules. J Comp Chem 1999;20:1091–1111.
28. Antosiewicz J, McCammon JA, Gilson MK. Prediction of pH-dependent Properties of Proteins. J Mol Biol 1994;238:415–436.
29. Nielson JE, Vriend G. Optimizing the hydrogen-bond network in Poisson-Boltzmann equation-based pK_a calculations. Proteins 2001;43:403–412.
30. Georgescu R, Alexov E, Gunner M. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. Biophysical Journal 2002;83:1731–1748.
31. Mongan J, Case D, McCammon JA. Constant pH Molecular Dynamics in Generalized Born Implicit Solvent. J Comp Comp Chem 2004;25:2038–2048.
32. Linderstrom-Lang K. On the ionization state of proteins. C R Trav Lab Carlsberg 1924;15:1–29.
33. Bashford D, Karplus M. Multiple site titration curves of proteins. J Phys Chem 1991;95:9556–9561.
34. Feig M, Onufriev A, Lee MS, Im W, Case DA, Brooks CL. Performance comparison of generalized Born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 2004;25(2):265–284.
35. Gilson MK. Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins. Proteins 1993;15:266–282.
36. Lee FS, Chu ZT, Warshel A. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the polaris and enzymix programs. J Comp Chem Feb 1993;14(2):161–185.
37. Nielsen J, McCammon J. Calculating pK_a values in enzyme active sites. Protein Sci Sep, 2003; 12(9):1894–1901.
38. van Vlijmen HWT, Shaefer M, Karplus M. Improving the accuracy of protein pK_a calculations: conformational averaging versus the average structure. Proteins 1998; 33:145–158.
39. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001; 44:400–417.
40. Onufriev A, Bashford D, and Case D. Modification of the generalized Born Model suitable for macromolecules. J. Phys. Chem. B 2000, 104:3712–3720.
41. Fetrow J, Siew N, and Skolnick J. Structure-based functional motif identifies a potential disulfide oxidoreductace active site in the serine/threonine protein phosphatase-1 subfamily. FASEB, J 1999, 13:1866–1874.
42. Fetrow J, Siew N, Gennaro JD, Martinez-Yamout M, Dyson J, and Skolnick J. Genomic-scale comparison of sequence- and structure-based methods of function prediction. Protein Sci 2001, 10:1005–1014.
43. Rothschild KJ. FTIR difference spectroscopy of bacteriorhodopsin: Toward a molecular model. J. Bioenerg. Biomembr. 1992, 24:147–167.
44. Haupts U, Tittor J, and Oesterhelt D. Closing in on bacteriorhodopsin: Progress in understanding the molecule. Annu. Rev. Biophys. Biomol. Struct. 1999, 28:367–399.
45. Lanyi JK. Molecular mechanism of ion transport in bacteriorhodopsin: insights from crystallographic, spectroscopic, kinetic, and mutational studies. J. Phys. Chem. B 2000, 104:11441–11448.
46. Gordon J, Myers J, Folta T, Heath LS, and Onufriev AH++: a server for estimating pK_as and adding missing hydrogens to macromolecules. Nucleic Acid Res 2005, 33:368–371.
47. Bashford D, and Gerwert K. Electrostatic calculations of the pK_a values of ionizable groups in bacteriorhodopsin. J. Mol. Biol. 1992, 224:473–486.
48. Tishmack PA, Bashford D, Harms E, and Van Etten RL. Use of 1h nmr spectroscopy and computer simulations to analyze histidine pK_a changes in a protein tyrosine phosphatase: Experimental and theoretical determination of electrostatic properties in a small protein. Biochemistry 1997, 36:11984–94.
49. Beroza P, and Case D. Including side chain flexibility in continuum electrostatic calculations of protein titration. J. Phys. Chem. 1996, 100(51):20156–20163.
50. Onufriev A, Case DA, and Ullmann GM. A Novel View of pH Titration in Biomolecules. Biochemistry 2001, 40(12):3413–3419.
51. Cormen T, Leiserson C, Rivest R, Stein C. Introduction to algorithms, second edition. Cambridge, MA: Massachusetts Institute of Technology; 2002.