Optimization in Computational Chemistry and Molecular Biology, pp. 1-2
C. A. Floudas and P. M. Pardalos, Editors

Exploring potential solvation sites of proteins by multistart local minimization

Sheldon Dennis, Carlos J. Camacho, and Sandor Vajda
Department of Biomedical Engineering
Boston University
44 Cummington Street, Boston MA 02215
vajda@bu.edu

Abstract

The thermodynamics of solvation is studied by exploring the local minima of a function that describes the free energy of water around a protein. In particular, we determine if the ordered water positions in the crystal become preferred solvent binding sites in solution. The free energy is obtained by determining the electrostatic field of the solvated protein from a continuum model, and then calculating the interactions between this field and a single water molecule. The local minima in the neighborhood of selected points are explored by two different approaches. The first is a simple mapping of the free energy on a grid. The resulting maps show that the "free energy pockets" around crystallographic water sites are clusters of local minima. The second approach is based on the classical simplex algorithm which is used in two different implementations, one with a penalty function and the other modified for constrained minimization, called the complex method. Both the simplex and the complex methods are much faster than mapping the free energy surface. The calculations are applied to T4 lysozyme with data available on the conservation of solvent binding sites in 18 crystallographycally independent molecules. Results show that almost all conserved sites and the majority of non-conserved sites are within 1.3 Å of local free energy minima. This is in sharp contrast to the behavior of randomly placed water molecules in the boundary layer which, on the average, must travel more than 3 Å to the nearest free energy minimum. Potential solvation sites, not filled by a water in the x-ray structure, were studied by local free energy minimizations, started from random points in the first water layer.

Keywords: Solvation free energy, local minimization, clusters of local minima, simplex method, complex method, test problem

1 Introduction

Proteins neither fold nor function without bound water molecules, and understanding solvation is clearly a central problem in the biophysics of macromolecules. The experimental

techniques that provide most information on the solvent structure around proteins are x-ray crystallography and NMR spectroscopy [1, 2]. In x-ray crystallography regions of extra electron density are interpreted as ordered water sites, resulting in about 200 water molecules in a typical high resolution protein structure [3]. NMR data reveal that in solution the water molecules around the protein are in rapid motion, including those that appear to be fixed in the x-ray structure. The exchange times are less than 500 ps with the exception of a few buried waters that may have residence times of up to 0.01 s [4, 5].

Since the electron density map derived from X-ray diffraction is averaged over a time scale measured in hours, discrete water density found in a crystal structure clearly does not indicate an actual water molecule at that position. It implies, however, that the potential of mean force has a local minimum; that is, the free energy of water must be lower than at all closely neighboring regions. As pointed out by Levitt and Park [3], if the free energy did not have a local minimum at that point, high electron density would not be found since, on average, water would be located relatively uniformly in that region.

The main computational tools used to study water behavior around proteins have been molecular dynamics and Monte Carlo simulations [6, 7, 8]. Such calculations confirmed that water molecules near the protein surface remain very mobile, with a diffusion coefficient decreased two- to four-fold relative to that of bulk water [8]. However, MD studies substantially differed in their evaluation of the influence of apolar, polar, and charged surface atoms on the mobility of the surrounding solvent [1, 7].

In this paper we search for preferred solvation sites by exploring local minima of a function that describes the free energy of water around the protein. The relationships between these preferred sites and the ordered water positions observed in the x-ray structure are also studied. We use a continuum model of electrostatics interactions [9, 10] in which the protein is represented by a low dielectric region containing discrete atomic charges at fixed positions, surrounded by a high dielectric medium representing the solvent, and calculate the electrostatic field of the solvated protein by solving the linearized Poisson-Boltzmann equation [9, 10]. The free energy of water is then calculated by translating and rotating a "probe" water molecule in the precalculated field. This approach removes the need for estimating the free energy change by averaging over a large number of trajectories as required in molecular dynamics, and reduces the problem to exploring the local minima of the free energy surface.

We focus on the following problems.

Problem 1: In order to determine if water positions in the x-ray structure are retained as preferred solvation sites in solution, we will explore the neighborhood of crystallographic water positions for local free energy minima.

Problem 2: In order to determine if there exist preferred solvation sites that are not in the vicinity of any crystallographic water site, we will start local minimization runs from a large number of points randomly placed in the first water layer.

In both problems we need robust local minimization algorithms that can explore the local minima in the vicinity of a starting point without jumping to other regions of the search space. Two approaches will be used. In the first, robustness is assured simply by calculating the free energy on a grid in the plane of the two Euler angles describing the

position of the water molecule, thus mapping the free energy surface in a neighborhood of selected points [11]. At each grid point minimization is still performed in the subspace of remaining variables. The resulting free energy maps show that the "free energy pockets" around crystallographic water sites are actually clusters of local minima, very close to each other and are separated by moderate free energy barriers.

While the free energy maps are informative, the method is far from efficient for finding local minima. Therefore we performed further calculations using the simplex algorithm [12], one of the most robust approaches to local minimization. Two versions of the method were used. In the first we employed a penalty function to represent excluded volume constraints, thereby converting the problem into an unconstrained one. The second version, referred to as the complex method, works directly with the constrained problem without a penalty function [13, 14] Both the simplex and the complex methods are much faster than mapping the free energy surface.

The calculations were applied to the x-ray structure of the T4 lysozyme that includes 137 ordered water sites. This system is particularly interesting because x-ray structures are available for 18 crystallographycally independent T4 lysozyme molecules, including the wildtype 4lzm and nine mutants [15]. The comparison of these structures provides valuable information on the conservation of water sites across different crystal forms of the same protein.

2 Methods

2.1 Conformational space

The free energy will be calculated for a "probe" water molecule translating and rotating around the protein. The position of each water molecule is given in a local coordinate system centered at some point $w_0 \in R^3$. In the free energy mapping we use spherical coordinates, i.e., the location of the water molecule is given in terms of the Euler angles ϕ and θ, and the radius r. Further three Euler angles, ϕ_w, θ_w, and ψ_w specify the orientation of the water. Thus, a point in this space is given by the vector $s = (r, \phi, \theta, \phi_w, \theta_w, \psi_w)^T$, and the search space is defined by $S = \{0 \leq r, 0^o \leq \phi \leq 360^o, 0^o \leq \theta \leq 180^o, 0^o \leq \phi_w, \psi_w \leq 360^o, 0^o \leq \theta_w \leq 180^o\}$. By contrast, the water position is given in Cartesian coordinates when the search is performed by the simplex or the complex method and a point is specified by the vector $s' = (\Delta x_w, \Delta y_w, \Delta z_w, \phi_w, \theta_w, \psi_w)^T$. In Problem 1, w_0 is a crystallographic water position, whereas in Problem 2 w_0 is a randomly selected point in the first water layer.

2.2 Free energy function

Let Φ denote the electrostatic field of a solvated protein. The free energy of a water molecule at position $s \in S$ is given by

$$\Delta G_{el}(s) = \sum_{i=1}^{3} \Phi(x_i) q_i \tag{1}$$

where x_1, x_2 and x_3 denote the positions of the atoms O, H_1, and H_2 in the water molecule, and q_1, q_2, and q_3 denote the corresponding atom-centered partial charges. We use the

TIP3 water model [16] which has rigid geometry, and hence for a given w_0 the vector $s \in S$ determines the atomic positions x_1, x_2 and x_3.

The electrostatic free energy is subject to steric constraints of the form

$$D_j < d_j, \;\; j = 1, ..., n \tag{2}$$

where d_j is the distance between the center of the water and the jth protein atom. The lower bound D_j is the sum of van der Waals radii of the jth protein atom and that of the water molecule, i.e., $D_j = r_j + r_w$.

To calculate the electrostatic field Φ, the linearized Poisson-Boltzmann equation is solved by a finite difference method as implemented in CONGEN [17]. The algorithm features adjustable rectangular grids, a uniform charging scheme that decreases the unfavorable grid energies, and smoothing algorithms that alleviate problems associated with discretization. The calculations were carried out using a 0.8 Å grid, with uniform charging, anti-aliasing, and 15-point harmonic smoothing. A 8 Å grid margin was maintained around the molecule. The dielectric constants of the protein and the solvent were set to 2 and 78, respectively, and the ionic strength was 0.05 M. Since the electrostatic field Φ is obtained only at the grid points, we employed a linear interpolation formula when calculating ΔG_{el} by Equation 1. Thus, with a precalculated field, free energy evaluation is extremely simple.

Before the free energy calculation, the x-ray structure of the protein has been refined by 200 steps of energy minimization, with the ordered water molecules included. We used version 19 of the Charmm potential [18] with polar hydrogens, and 20 Kcal/mol/Å ^{2} harmonic constraints on the positions of non-hydrogen atoms. These calculations placed the polar hydrogens and created a plausible hydrogen-bonding network between protein and water. The RMS shift of the water molecules due to the minimization was 0.1 Å.

2.3 Free energy maps

In order to understand the properties of the function defined by Equation 1 we first mapped the free energy surface in the vicinity of crystallographic solvent sites (Problem 1). Let w_0 denote the position of an ordered site. In the spherical coordinate system placed at w_0 we consider the lines defined by Euler angles (ϕ_i, θ_i), where $\phi_i = 0^o, 18^o, ...342^o$, and $\theta_i = 0^o, 18^o, ..., 180^o$. The free energy is minimized along each of these lines in terms of the remaining variables $(r, \phi_w, \theta_w, \psi_w)$ subject to the steric constraints given by Equation 2.

The constrained minimization problem is solved using a penalty function approach, i.e., by the unconstrained minimization of the extended target function

$$Q = \Delta G_{el} + CV_{exc} \tag{3}$$

where V_{exc} is an excluded volume penalty function defined by

$$V_{exc} = \begin{cases} \sum_{j=1}^{n} (D_j - d_j)^2 & \text{if } d_j < D_j; \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

and C is a weighting coefficient. As we will show, using $C = 10^5$ the atomic overlaps after the convergence of the minimization are so small that the ΔG_{el} term is not significantly affected. Notice that using the extended potential any steric overlap will show up as a sudden increase on the free energy maps.

For each (ϕ_i, θ_i) our goal is to find the local free energy minimum closest to the original water position w_0 along the line defined by (ϕ_i, θ_i), i.e., the smallest displacement r at which a local minimum occurs. Therefore we select $r = 0$ as the initial displacement. However, we also need to choose starting values for the variables $(\phi_w, \theta_w, \psi_w)$ that describe the orientation of the "probe" water molecule. As may be expected in a nonlinear problem and is confirmed by our calculations, for a given (ϕ_i, θ_i) and initial point $r = 0$ it is possible that, for some initial values of $(\phi_w, \theta_w, \psi_w)$, the Powell method does not find the local minimum corresponding to the smallest displacement r. In some cases the search may terminate early, e.g., due to being restricted to a subspace of the four-dimensional search space; in others the procedure may jump to a more distant local minimum. These artifacts can be easily identified by repeating the minimization for each (ϕ_i, θ_i) 30 times with different random orientations of the probe as the starting state. Since early termination or jumping over a local minimum along the line defined by (ϕ_i, θ_i) are relatively rare, the majority of the 30 runs yields very similar displacement values, well distinguishable from the few runs that end up in substantially different points. For each (ϕ_i, θ_i) we consider only the highly populated cluster of final water displacements, and choose the lowest ΔG_{el} value within this cluster as the free energy minimum along the line defined by the Euler angles (ϕ_i, θ_i).

Restricting consideration to the above minima the free energy surface can be visualized using two maps. The first map shows the minimum free energy as a function of ϕ_i and θ_i. The second map shows the displacement r at which the minimum occurs. The two maps together show if there is a free energy pocket in the vicinity of the crystallographic water site, and how far this pocket extends along each direction before the free energy starts to increase. We note that the analysis of the final maps provides an independent method to test if a local minimum found by the Powell method is the closest one to the origin along the line (ϕ_i, θ_i). In fact, an early termination or jump shows up as a discontinuity at the corresponding grid point on the displacement map, and most frequently also on the free energy map. As we will further discuss, selecting the dominant cluster from the 30 runs for each (ϕ_i, θ_i) assured the continuity of both maps for almost all water molecules.

2.4 Minimization by the simplex method

The classical simplex method is for unconstrained minimization, and will be used with the extended function defined by Equation 3. The algorithm proceeds as follows [12].

1. Select a starting simplex represented by $k = n + 1$ vertices $x^{(1)}, x^{(2)}, \ldots, x^{(k)}$. The method actually starts with only one feasible point $x^{(1)}$, and the remaining $k - 1$ points are found one at a time by random selection in the neighborhood of $x^{(1)}$.

2. Evaluate the target function Q at the k vertices and select the worst point x^{max} such that $f(x^i) \leq f(x^{max})$ for $i = 1, 2, \ldots, k$, and the best point x^{min} such that $f(x^i) \geq f(x^{min})$ for $i = 1, 2, \ldots, k$,

3. Calculate the the centroid of the simplex by

$$\overline{x} = [\sum_{i=1}^{k} x^{(i)} - x^{max}]/n \tag{5}$$

Notice that the centroid excludes the worst point x^{max}.

4. Calculate (1) the reflection point x^* with $x^* = (1+\alpha)\overline{x} - \alpha x^{max}$, where $\alpha > 0$, (2) the expansion point x^{**} with $x^{**} = \gamma x^* + (1-\gamma)\overline{x}$, where $\gamma > 1.0$, and (3) the contraction point x^{***} with $x^{***} = \beta x^{max} + (1-\beta)\overline{x}$, where $0 \leq \beta \leq 1.0$

5. Evaluate the function Q at the reflection point. If the reflection point is better than the best point, evaluate Q at the expansion point x^{**}. If the expansion point is better than the reflection point, replace the worst point in the simplex by the expansion point, otherwise replace it by the reflection point.

6. If the reflection point is better than the worst point, replace the worst point by the reflection point in the simplex. Otherwise evaluate Q at the contraction point. If the contraction point is better than the worst point, then replace the worst point by the contraction point in the simplex; otherwise reduce the size of the simplex leaving only the best point in place [12].

7. Terminate the iteration if the norm in the correction of the centroid is smaller than a threshold ϵ.

The algorithm has great versatility in adopting the simplex to the local free energy landscape. It will elongate and take a larger step if it can do so, it will change direction on encountering free energy barriers at an angle, and it will contract in the neighborhood of a minimum. However, for our purposes the most important feature of the method is its robustness which can be further increased by defining an upper bound on any side of the simplex, which will assure that the method will explore nearby local minima and will not jump to a far point of the conformational space. Although we may need to evaluate the function at more points than for a method with a superlinear convergence rate, the steps made by the simplex generally provide useful information on the form of the surface.

The simplex algorithm obviously provides only a local minimization tool. As we will further discuss, in the present application we want to explore the local minima in the neighborhood of a starting point. This will be accomplished by performing 30 minimization runs with randomly generated simplexes placed around the given starting state. The points representing the results of the 30 minimizations are clustered in the Cartesian space, i.e., differences in the rotational coordinates ϕ_w, θ_w, and ψ_w are ignored. We selected a simple clustering algorithm (see http://mvhs1.mbhs.edu/mvhsproj/projects/clustering/algorithm2.html). The original method introduces an appropriate number of clusters such that the distances within each cluster are smaller than the half-distance between any two clusters, where the inter-cluster distance is defined as the distance between the hubs of the corresponding clusters. This algorithm was modified by introducing a lower bound L on the inter-cluster distances, and thus any two clusters with a distance below L are concatenated into a single cluster. As we will discuss, this lower bound accounts for the fact that water positions in two different crystal forms of the same protein have been clustered into a single solvation site if they were closer than 1.2 Å to each other [15, 20].

2.5 Minimization by the complex method

The complex method is a straightforward extension of the simplex method ([14], p. 292) for solving constrained problems without introducing a penalty term. The method can handle only inequality constraints. The complex method is essentially a simplex method with the additional condition that any vertex must be a feasible point. Thus, every time a trial point

$x^{(j)}$ is generated, we find whether it satisfies all the constraints. If it violates any of the constraints, the point is moved half way toward the centroid of the already accepted points. This reduction in step size is continued until a feasible point is found. We will ultimately be able to find a feasible point by this procedure provided the feasible region is convex.

For comparison we tried to keep decision making in simplex and complex algorithms as similar as possible, but there are some essential differences. In the simplex method we use $k = n + 1$ vertices as suggested by Nelder and Mead [12]. A larger k leads to early convergence and generally causes the method to miss the extreme point being searched for. By contrast, for the complex method Box [13] recommended a value $k \approx 2n$, and we used $k = 2n$. In fact, if k is not sufficiently large, the complex tends to collapse and flatten along the first constraint boundary encountered. The other difference is in the selection of the starting simplex. In the case of the simplex method the only requirement is to "fill" the entire space, i.e., to avoid the flattening of the simplex to a subspace. By contrast, the complex method requires points that are feasible. As we described, feasibility can be assured by reducing the simplex if the feasible region is convex and at least one feasible point is already available. The first feasible point is placed by a randomized trial-and-error procedure. As in the case of the simplex method, we perform 30 minimization runs with different initial simplexes placed in the vicinity of the given starting point, and cluster the solutions to obtain information on the nearby local minima.

2.6 Application

We use the 1.7 Å resolution x-ray structure of the T4 lysozyme (PDB code 4lzm) that includes 139 water molecules. We placed the polar hydrogens and removed potential steric overlaps by 200 steps of energy minimization using version 19 of the Charmm potential [18] and 20 Kcal/mol/Å^2 harmonic constraints on the positions of non-hydrogen atoms. The refinement of the crystal structure results in an RMS shift of 0.1 Å.

T4 lysozyme is studied because the x-ray structure is available for a number of mutants that differ from the wildtype only at one or two positions [15]. Although the mutations affect the structure in the immediate vicinity of the amino acid substitution, the rest of the protein remains essentially unchanged. In some cases some "hinge-bending" motion occurs, but it can be corrected by bringing the rigid body fragments into a reference frame. The overlap of 18 x-ray structures results in a total of 1675 water molecules. Zhang and Matthews [15] clustered these waters by taking each water in turn, and counting how many other water molecules occurred within 1.2 Å, forming a total of 139 clusters. The analysis of these clusters provided information on the conservation of solvation sites. In particular, 40 sites that were occupied in at least 7 of the 18 molecules have been defined as conserved water positions.

3 Results and Discussion

In this paper we focus on the simplex and complex methods. A summary of free energy mapping results is given, with details reported elsewhere [21]. As we will show, the maps help to understand the optimization problem and the behavior of the algorithms.

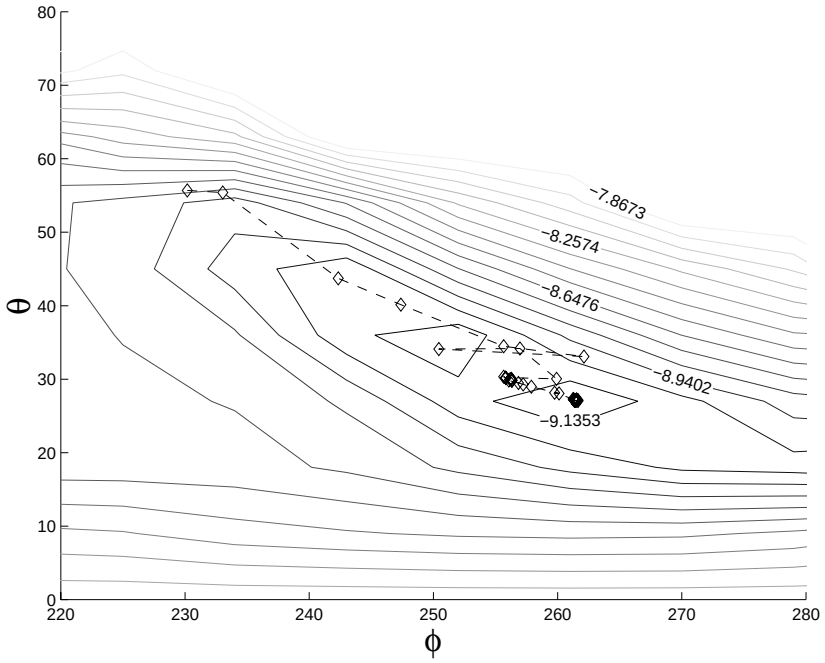

Figure 1: Contour lines of the minimum free energy surface in site 311 as a function of the Euler angles ϕ and θ. The dashed line shows the trajectory in a simplex minimization.

3.1 Free energy maps

Free energy and displacement maps have been constructed for the neighborhoods of crystallographic water sites in the x-ray structure 4lzm.

As an example, Figures 1 and 2 show such maps for water site No. 311. Figure 1 shows the minimum free energy as a function of the Euler angles ϕ and θ, i.e., the lowest free energy values attained along the lines defined by (ϕ, θ) as we move outward from the center of the local coordinate system. Figure 2 shows the displacement r at which this minimum is attained.

The free energy and displacement maps together provide substantial information on the local behavior of the free energy function. Based on the free energy maps we conclude that the free energy pockets discussed by Levitt and Park [3] are actually clusters of local minima, generally separated from other clusters by relatively high free energy barriers. Due to these energy barriers most clusters are well defined, and hence for each cluster we can identify the local minimum with the lowest free energy and determine its displacement r from the crystallographic water position. However, since there may exist several local minima with similar energies, we frequently study all minima within the cluster rather than focusing on the lowest energy one.

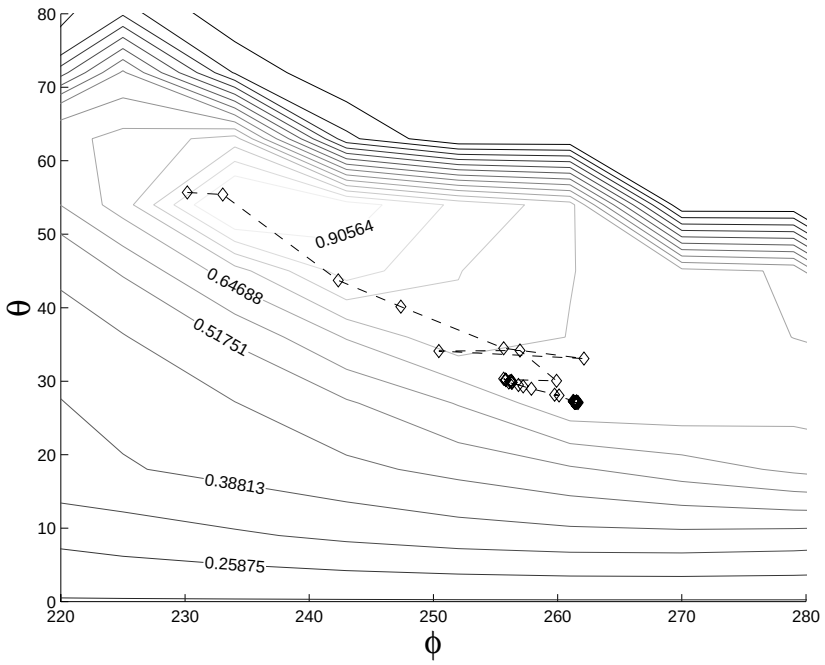

Figure 2: Displacement r at the free energy minima shown in Figure 1, including the pathway of the centroid in the simplex minimization.

Although the particular geometries differ, the site around water 311 is typical of most crystallographic water sites. Figure 1 shows only the region containing the two lowest energy minima, but the free energy pocket around this site includes two other minima that are beyond the boundaries of the plot. As the displacement map in Figure 2 shows, the cluster is well defined, outward moves from the crystallographic position may reduce the free energy for some distance, but after less then 1 Å displacement the free energy invariably starts to increase. In many cases the border of the pocket is very sharp because even a very small move along the line defined by the corresponding ϕ and θ pair steeply increases the free energy. Such steep increase is generally due to steric overlaps, indicating that the site is close to the protein surface. For water 311 the maximum displacement occurs at $\phi = 235^o$ and $\theta = 55^o$, and it is close to 1 Å. The lowest free energy value, -9.3 kcal/mol, is at $\phi = 261^o$ and $\theta = 29^o$ (Figure 1), and it is attained with slightly less than 0.8 Å displacement (Figure 2).

We attempted to construct free energy and displacement maps for all the 139 crystallographic water sites in 4lzm. However, in seven cases the maps showed sudden changes in the free energy at some points of the (ϕ, θ) plane.This may signal an inherent discontinuity of the free energy function due to the instability of the water molecule at the particular sites when the protein is in solution rather than in a crystal. However, as we described in the method, it is also possible that our procedure of performing 30 local minimizations at (ϕ_i, θ_i) and then selecting the most populated cluster is unable to identify the minimum

closest to the original crystallographic site. Therefore we increased the number of minimization runs to 60, but were unable to remove the discontinuities. As we will discuss, the seven water sites were re-examined by the simplex method, and were shown to be inherent discontinuities.

Of the 132 remaining sites, 100 are within 1.3 Å of the closest local free energy minimum. In terms of displacement statistics, there is a slight difference between conserved and non-conserved water positions. The closest minimum is within 1.3 Å for the majority of conserved sites (31 out of the 36, or 86.1%). In the case of non-conserved sites, this fraction is 71.8%. However, in spite of the substantial shifts in the position of a few non-conserved waters, a comparison of displacement distributions for conserved and non-conserved sites does not show a significant difference. Similarly, there is no significant difference between the free energy distributions, although a few buried water molecules have very low free energies (≤ -12 kcal/mol).

The finding that most crystallographic sites are within 1.3 Å of a local minimum is in sharp contrast to the behavior of water molecules randomly placed in the first water layer. Such molecules, on the average, must move 3.2 Å to reach the nearest local minimum. Thus, we conclude that ordered water sites in the x-ray structure are at least partially due to favorable electrostatic interactions between the protein and the water, and the majority of such positions remain preferred solvation sites when the protein is in solution.

3.2 Search by simplex and complex methods: crystallographic sites

The free energy maps show the local minima in the neighborhood of 132 crystallographic water sites. To test the simplex and complex methods, we explored the same regions by performing 30 minimization runs with randomly selected initial simplexes around each site. As mentioned in Methods, the search is in the space defined by the vectors $s' = (\Delta x_w, \Delta y_w, \Delta z_w, \phi_w, \theta_w, \psi_w)^T$. In the translational subspace, the vertices of the initial simplex have been obtained by random displacements between -1 Å and 1 Å along each coordinate axis. The rotational coordinates ϕ_w and ψ_w of the vertices have been randomly selected in the 0^o to 360^o interval and θ_w from the 0^o to 180^ointerval.

Figures 1 and 2 also show the trajectory of the centroid in a minimization by the simplex method, superimposed on the free energy and displacement maps, respectively. The trajectory shown contains 350 simplex iterations. As described in the method, for each water site we perform 30 minimization runs from different initial points, and cluster the solutions. For water 311, 24 of the 30 solutions form a single cluster (are within 1.2 Å to each other), and the remaining 6 points distribute among three other clusters, indicating three further local minima. Comparison with the free energy map shows that the most populated cluster is also the lowest free energy local minimum.

Figures 3 and 4 show the trajectory of the centroid in a minimization by the complex method. Since the algorithm includes a search for a feasible initial simplex in the $k = 2n = 12$-dimensional space, the initial position of the centroid generally differs from the one in the simplex method, even when the same seed is used in the random number generator. As shown by the relatively high energy, for the particular run in Figure 3 the initial simplex is in a region that is close to the protein but is still feasible (i.e., the extended target function is below -7 kcal/mol, indicating the lack of any steric overlap). Without a penalty term,

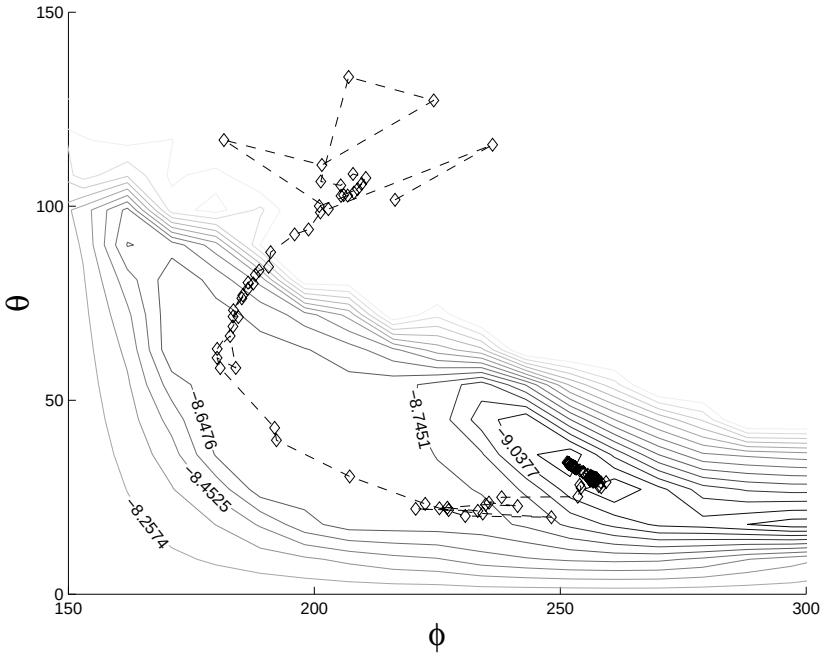

Figure 3: The pathway of the centroid of the simplex in a minimization by the complex method in site 311. The contour lines represent the minimum free energy surface shown in Figure 1.

the electrostatic component ΔG_{el} in this region changes slowly, and thus the simplex moves around before starting toward the minima in the free energy pocket. The trajectory of the centroid shown is actually the result of over 1,500 iterations.

Table 1 lists the average number of function evaluations and the average CPU time for the three different methods, the latter for runs on a single SiliconGraphics R10000 processor. For the grid-based approach both numbers are shown for the 30 minimization runs by the Powell method at a given (ϕ_i, θ_i) in the space of the four variables $(r, \phi_w, \theta_w, \psi_w)$. Notice that selecting the most populated cluster provides the free energy and displacement values at (ϕ_i, θ_i), i.e., a single grid point of the free energy and displacement maps. Since with the selected grid density each map consists of 220 grid points, the total average CPU time required for the construction of one map is 27,280 s, or about 7.6 hrs on a single R-10000 processor.

For the simplex and complex methods Table 1 shows the (average) total number of function evaluations and the (average) total CPU time in 30 minimization runs, starting with different initial simplexes. In contrast to the 30 minimizations in the grid-based approach that yield a single point of the free energy map, the 30 runs explore an entire cluster of local minima (see below). As shown in the Table, the general relationship between simplex and complex trajectories is well represented by Figures 1 and 3; i.e., the complex method, on the average, requires about five times as many iterations as the simplex method. In spite

of this large difference in the number of function evaluations, the CPU times differ only by about 16%. This apparent contradiction is due to the extreme simplicity of function evaluation in this particular case, which is essentially interpolating in a table. Therefore, the CPU times are determined by the computational overhead rather than by the number of function evaluation, and the former seems to be very similar for simplex and complex methods. Both are much more efficient than mapping the free energy surface. For example, exploring a water site with the simplex method, on the average, requires 807 s and thus about 13 minutes instead of the 7.6 hrs CPU time for constructing a free energy map.

Table 1. *Performance characteristics of the three methods*

	Method		
	Grid-based	Simplex	Complex
Number of function evaluations	5153 ± 5112	6063 ± 360	28428 ± 2559
CPU time, s	124 ± 16	807 ± 109	944 ± 134

Why does the simplex method require much fewer function evaluations than the complex method? We recall that the simplex and complex methods work with $k = n + 1$ and $k = 2n$ vertices, respectively. We have evidence that using fewer vertices in the simplex method has a favorable effect, because increasing the number of vertices without any other change in the simplex algorithm increases the number of function evaluations. However, the better performance of the simplex method is mainly due to the use of the penalty function given by Equation 2 that facilitates the elimination of steric overlaps. Since the effect of the penalty term on the position of the minima can be neglected, particularly at the limited resolution of the grid, in the application considered here the simplex method with a penalty function is superior to the complex method that accounts for the constraints in a binary fashion, without any concept of direction.

For each water site, the results of the 30 runs are clustered as described in the Methods, using 1.3 Å as the lower bound L on the cluster size. Each cluster essentially represents a set of points that are within the region of attraction of a local minimum, and is labeled by the name of the atom closest to the hub of the cluster. By comparing the lists of local minima to the free energy maps we concluded the 30 runs for each site were sufficient to find the lowest minimum in each free energy pocket as one of the clusters, but not necessarily the lowest energy ones: for the 139 crystallographic sites, the most populated cluster is also the lowest energy one in 92 cases (66%). Comparison to the free energy maps shows that it is more meaningful to characterize the position of the free energy "pocket" (i.e., the cluster of local minima separated from other clusters by free energy barriers) by the position of the most populated cluster rather than by the position of the lowest energy cluster. The explanation is that in spite of the simplicity of the simplex method, some trajectories may end up in a different cluster, particularly if the free energy of the particular "pocket" is much

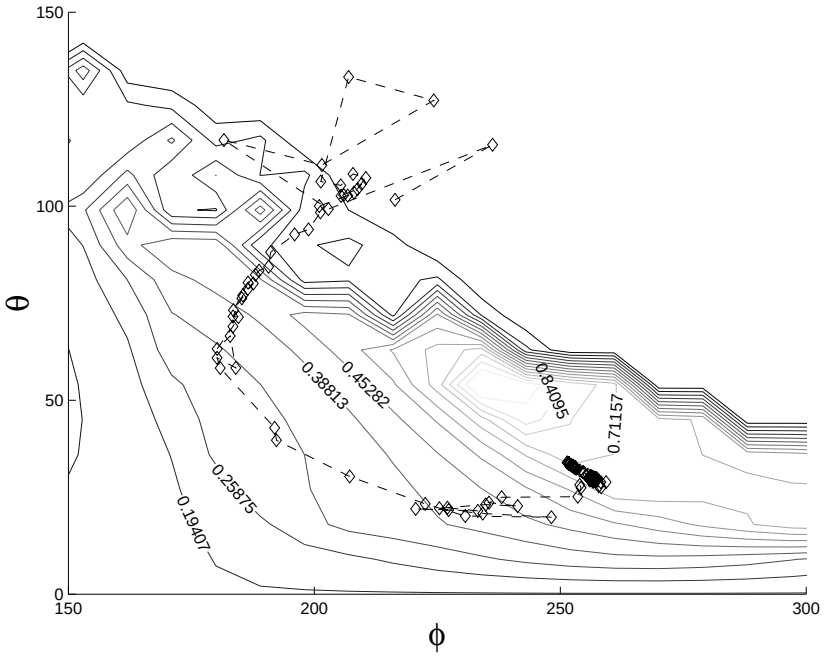

Figure 4: Displacement r at the free energy minima shown in Figure 3, including the pathway of the centroid in the complex minimization.

higher than the free energy of a neighboring pocket. However, the "jump" to a different cluster happens only in a small fraction of the trajectories, and hence the displacement of the most populated cluster is a more robust characterization of the distance between the original water position and the free energy pocket. Indeed, the most populated cluster is at the local minimum closest to the original site in 105 of the 139 crystallographic sites (75%). Furthermore, for the 36 conserved water sites that are generally better defined than the non-conserved sites, the most populated cluster is also the closest one in 29 cases (81%). In 4 of the remaining 7 sites the most populated cluster contacts the side chain that interacts with crystallographic water in the x-ray structure.

Using the simplex and complex methods we were able to explore the seven water sites for which no meaningful free energy maps have been obtained due to the sudden changes. It turns out that all corresponding free energy pockets are very shallow, resulting in five to eight clusters. In two cases (water sites 205 and 209) some trajectories end up close to the crystallographic water positions, but at very high energies (-1.95 kcal/mol and -2.07 kcal/mol, respectively), although for water 209 other water positions (230 and 337) are also obtained. For the other five sites (220, 233, 234, 255 and 260) none of the trajectories remains close to the crystallographic positions. Since all the seven sites discussed here are non-conserved, the weak energetics do not contradict the experimental evidence. However, it remains an open question as to why these water molecules are ordered in the x-ray structure.

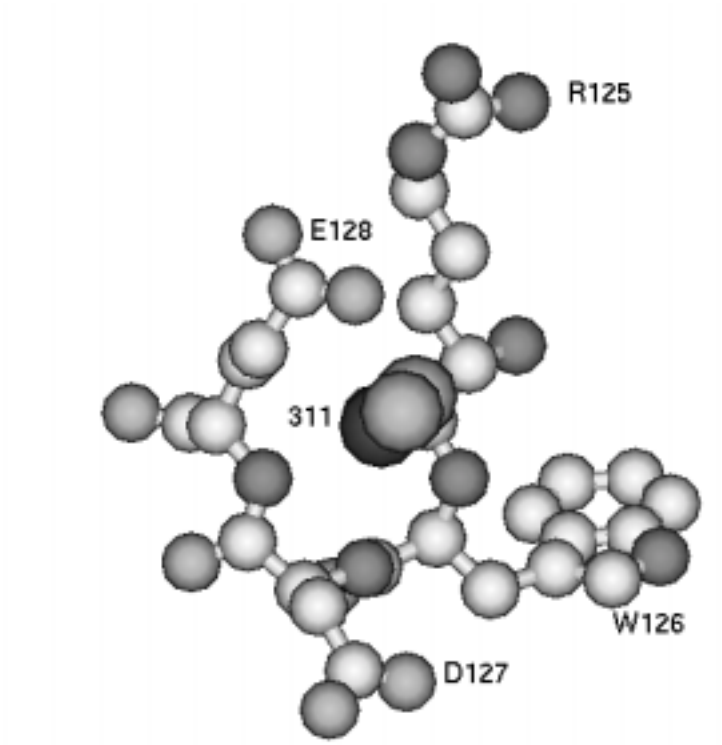

Figure 5: The two most populated local minima (large shaded spheres) in the simplex search from water position 311 (large dark sphere). Residues 125 through 128 of the T4 lysozyme are also shown using shades as follows: carbons lightly shaded, oxygens darker, and nitrogens dark.

For the 139 water sites the calculations yield 570 clusters (i.e., 4.1 clusters per site), 203 of which (40%) correspond to water molecules For 130 of the 139 sites, the atom closest to one of the clusters is an atom of the original water in its crystallographic position, with hydrogens placed by Charmm [18]. The corresponding clusters are generally well populated. This result is in complete agreement with the results of the free energy mapping, and confirms that the majority of crystallographic positions become preferred solvation sites in solution.

The 203 water clusters for the 139 sites means that for a number of water molecules some of the 30 trajectories end up in another water site, i.e., the crystallographic sites are not necessarily separated by high energy barriers. In addition, there are 367 non-water clusters. Most of these are close to side chains that interact with the particular water molecule in the x-ray structure. However, the additional minima are far enough from the crystallographic water sites to be regarded as separate clusters. In fact, high resolution protein structures reveal that a charged side chain, particularly Glu or Asp, can simultaneously contact up to five ordered water sites. The additional clusters correspond to such potential solvation sites that in the given x-ray structure are not filled with a well determined water. As we will see, most of these sites are on the "other side" of side chains that already contact at least one ordered water.

For example, Figure 5 shows the two most populated local minima (large gray spheres) when starting minimizations from water position 311 (large black sphere). Residues 125

through 128 of the T4 lysozyme are also shown. Notice that the site is surrounded by the charged side chains of Asp-127 and Glu-128, as well as by Arg 125 that is a bit farther from the water position. Rather than shifting to any of these side chains, water 311 is seen in a free energy pocket that is cooperatively determined by the three, and by additional charges on nearby backbone atoms. Free energy minimizations robustly identify this region of the free energy, but they also show that the region contains several local minima and thus several potential water positions, two of which are shown in Figure 5. Since these minima are close to each other and are separated by moderate free energy barriers, water molecules can shift from one to another.

As we have discussed, the most populated cluster is generally a useful characterization of the entire free energy pocket and hence we also collected statistics by restricting consideration to these clusters. The corresponding free energy was negative for all sites. The closest atom to the cluster was that of a water molecule for 103 of the 139 sites, but in two cases the trajectory actually ended close to a different water molecule, not the one in the original site. In 18 cases the populated cluster was close to a side chain that contacts a water molecule in the x-ray structure. This means that the minimized water orientation relative to the side chain differs from the one observed in the crystal. In 13 cases a backbone atom is the one closest to the populated minimum. The free energy of such water molecules ranges from -6.1 to -3.8 kcal/mol. Although these interactions are relatively weak (the free energy of some water molecules can be as low as -13 kcal/mol), the mean free energy for backbone-contacting water molecules is not significantly higher than the mean free energy for all ordered sites. In fact, some of the crystallographic water sites interact with the backbone.

There are 5 cases in which the minimized water contacts a side chain that has not been observed to contact an ordered water in the x-ray structure. Most frequently this situation occurs for charged or polar side chains that could interact favorably with water molecules, but are too ill-defined to be seen in the x-ray structure. In fact, Thanki et al. [22] observed that the highly mobile Lys side chains have no preferred orientation for water contacts. The same applies to most Tyr and Trp side chains.

We note that the initial simplex size, particularly in the translational subspace, is an important parameter that controls how well the local minima in a neighborhood are explored. If the size is too small, all runs remain in the same local minimum, whereas a simplex that is too large may move over to a different site. With the current selection at least 90% of trajectories remained in the free energy pocket, and essentially all local minima, observable on the free energy maps, have been found. We also found additional clusters corresponding to local minima that were not seen on the maps, most likely due to their limited resolution.

3.3 Search by simplex and complex methods: random positions

Although the results just described already provide some information on the potential water positions (i.e., free energy minima that do not correspond to crystallographic water sites), we randomly selected 100 water positions in the first solvation layer and explored the local minima in their neighborhoods by the simplex method. The hundred runs results in 414 clusters, i.e., the number of clusters per residue is 4.1, exactly the same as in the minimizations around crystallographic sites. Only 29 of these clusters are closest to a water site.

This is 7% of the total, much less than around the ordered sites.

Restricting consideration to the most populated clusters we drop 8 minimization runs in which the final free energy is positive. 14 clusters are closest to an ordered water site. Although in the remaining 78 trials the most populated clusters are closer to protein atoms than to a crystallographic water, 36 of these interact with side chains that also contact at least one ordered water. There are 26 trials in which the dominant cluster moves to a backbone atom. For some of these sites the water-protein interaction is relatively weak, but the mean does not significantly differ from the mean calculated for all sites.

In the remaining 16 trials we find side chains that are not seen to contact an ordered water in the x-ray structure. These include four Lys residues, Lys-16, Lys-43 (twice), Lys-60, and Lys-162. Based on the thermal factors, all these side chains are very mobile in the crystal, and water molecules associated with them would not be seen. There are two Asn side chains, Asn-2 and Asn-81, that interact weakly with water. According to our results, the side chains of Asp 72 and Asp 92 must also be preferred solvation sites, and we do not fully understand why no ordered water is seen around them in the x-ray structure.

4 Conclusions

The structural and functional importance of water associated with proteins is well known [1, 2]. While x-ray crystallography and NMR techniques provide a wealth of structural data, they leave open important questions concerning the origin of preferred water sites, and the relationship between water positions in crystal structures and in solution. On the basis of the experimental data and the somewhat incomplete theoretical studies, Levitt and Park [3] formulated a number of assumptions. To explain the origin of an ordered water they concluded that the potential of mean force at that point must have a local minimum, i.e., the free energy of a water at closely neighboring regions (within say 2-3 A) must be relatively high, forming a free energy "pocket".

In this paper we studied the above free energy "pockets" by performing local minimization runs in the neighborhoods of interest, and then clustering the obtained local minima. The free energy was calculated for an explicit water molecule ("probe" molecule), interacting with the electrostatic field of the solvated protein. Using a precalculated field this approach results in a very simple free energy function. The computational problem is to find the local minima that are closest to a given point in space. However, accounting for both translation and rotation of the "probe" water, the search is in the space of six variables: the Euler angles ϕ and θ, the center-to-center distance r, and the Euler angles ϕ_w, θ_w, and ψ_w that specify the orientation of the water. The minimization is subject to steric constraints that prevent water-protein overlaps.

We used two different approaches to find the local minima in the neighborhood of selected points. The first employs a penalty functions and attempts the minimization by Powell's method. We have found that this method lacked the necessary robustness, i.e., depending on the initial values for the rotational variables ϕ_w, θ_w, and ψ_w, the algorithm frequently converged to distant local minima, ignoring others close to the initial point. We improved the performance by reducing the dimensionality of the problem, i.e., by converting it into a grid-based search for the variables ϕ and θ, and performing repeated local minimizations in the space of the remaining variables r, ϕ_w, θ_w, and ψ_w. Thus, in terms of

ϕ and θ, this is a mapping of the free energy on a grid. The resulting maps showed that the free energy "pockets" around crystallographic water sites are actually clusters of local minima, and the different clusters are generally but not always separated by relatively high free energy barriers.

The second approach was based on the classical simplex algorithm which was used in two different implementations, one with a penalty function and the other modified for constrained minimization, called the complex method. Both the simplex and the complex methods were much faster than mapping the free energy surface. The methods have been used with multiple runs to explore the various local minima in the free energy "pockets". In particular, we have selected 30 random initial simplexes, performed the searches, and clustered the solutions. Each cluster essentially represented a set of points that were within the region of attraction of a local minimum. As we have shown, the positions of the most populated clusters provided the best key to describe the free energy pocket.

The calculations were applied to T4 lysozyme with data available on the conservation of solvent binding sites in 18 crystallographycally independent molecules. Results show that the majority of crystallographic water sites are within 1.3 Å of local free energy minima. This is in sharp contrast to the behavior of randomly placed water molecules in the boundary layer which, on the average, must travel more than 3 Å to the nearest free energy minimum. Potential solvation sites, not filled by a water in the x-ray structure, were studied by local free energy minimizations, started from random points in the first water layer, and generally found good agreement with the crystallographic site. There were some false positives, i.e., local minima that were not seen in the crystal structure as ordered water sites. However, the dominant majority of these false positives were not very wrong, because they were simply on the "other side" of side chains that contacted one or more water molecules anyway, but not in the same orientation.

Acknowledgments: This research has been supported by grant DBI-9630188 from the National Science Foundation and by grant DE-F602-96ER62263 from the Department of Energy.

References

[1] Karplus, P.A., Faerman, C.H. (1994), "Ordered water in macromolecular structure," *Curr. Opinion Struct. Biol. Vol. 4*, 770-776.

[2] Teeter, M.M. (1991), "Water-protein interactions: Theory and experiment," *Annu. Rev. Biophys. Biophys. Chem. Vol. 20*, 577-600.

[3] Levitt, M., Park, B. (1993), "Water: now you see it, now you don't," *Structure, Vol. 1*, 223-226.

[4] Otting, G., Liepinsh, E., Wuthrich, K. (1991), "Protein hydration in aqueous solution," *Science Vol. 254*, 974-980.

[5] Belton, P.S. (1994), "NMR studies of protein hydration," *Prog. Biophys. Mol. Biol. Vol. 61*, 61-79.

[6] Levitt, M., Sharon, R. (1988), "Accurate simulation of protein dynamics in solution," *Proc. Natl. Acad. Sci. USA. Vol. 85*, 7557-7561.

[7] Brunne, R.M., Liepinsh. E., Otting, G., Wuthrich, K,, van Gunsteren, W.F. (1993), "Hydration of proteins. A comparison of experimental residence times of water molecules solvating the bovine pancreatic trypsin inhibitor with theoretical model calculations," *J. Mol. Biol. Vol 231*, 1040-1048.

[8] Makarov, V.A., Feig, M., Andrews, B.K, Pettitt, B.M. (1998), "Diffusion of solvent around biomolecular solutes: a molecular dynamics simulation study," *Biophys. J. Vol. 75*, 150-158.

[9] Gilson, M.K., Honig, B. (1988), "Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis," *Proteins Vol. 4*, 7-18.

[10] Honig, B., Nicholls, A. (1995), "Classical electrostatics in biology and chemistry," *Science Vol. 268*, 1144-1149.

[11] Camacho, C.J., Weng, Z., Vajda, S.,and DeLisi, C. (1999), "Free energy landscapes of encounter complexes in protein-protein association," *Biophys. J. Vol 76*, 1176-1178.

[12] Nelder, J.A., Mead, R. (1964) " A simplex method for function minimization," *Computer J. Vol. 7*, 308-313.

[13] Box, M.J. (1965) "A new method of constrained optimization and a comparison with other methods," *Computer J. Vol. 8*, 42-52.

[14] Rao S.S. (1978), *Optimization. Theory and Applications*, John Wiley and Sons, New York.

[15] Zhang, X.J., Matthews, J.W. (1994), "Conservation of solvent-binding sites in 10 crystals of T4 lysozyme," *Prot. Science Vol 3*, 1031-1039.

[16] Jorgensen, W.L, Chandrasekhar, J., Madura, J.D. (1983), "Comparison of simple potential functions for simulating liquid water," *J. Chem .Phys. Vol. 79*, 926-935.

[17] Bruccoleri, R.E. (1993), "Grid positioning independence and the reduction of self–energy in the solution of the Poisson–Boltzmann equation," *J. Comp. Chem. Vol. 14*, 1417-1422.

[18] Brooks, B.R., Bruccoleri, R.E., Olafson, B., States, D.J., Swaminathan, S., Karplus, M. (1983), "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *J. Comp. Chem. Vol. 4*, 197-214.

[19] Press, W., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T. (1990), *Numerical Recipes*, Cambridge University Press, Cambridge.

[20] Sanschagrin, P.C., Kuhn, L.A.(1998), "Cluster analysis of concensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity," *Prot. Science Vol. 7*, 2054-2064.

[21] Dennis, S., Camacho, C., Vajda, S. (1999), "A continuum electrostatic analysis of preferred solvation sites around proteins in solution, " *Proteins*, in press.

[22] Thanki, N., Thornton, J.M., Goodfellow, J.M. (1988), "Distribution of water around amino acid residues in proteins," *J. Mol. Biol. Vol. 202*, 637-657.