Proteins in Mixed Solvents: A Molecular-Level Perspective

Brian M. Baynes and Bernhardt L. Trout*

Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received: August 12, 2003; In Final Form: October 3, 2003

We present a statistical mechanical approach for quantifying thermodynamic properties of proteins in mixed solvents. This approach, based on molecular dynamics simulations which incorporate all atom models and the theory of preferential binding, allows us to compute transfer free energies with experimental accuracy and does not incorporate any adjustable parameters. Specifically, we applied our approach to the model proteins RNase A and T1 and the solvent components water, glycerol, and urea. We found that the observed differences in the binding of glycerol and urea to RNase T1 and A are predominantly a consequence of density differences in the first coordination shell of the protein with the cosolvents, but the second solvation shell also contributes to the overall binding coefficients. Our approach allows us to determine the contributions of individual sites on a protein to preferential binding. One conclusion from determining these contributions is that hydrophobic amino acids in RNase T1 tend to bind less water and more cosolvent molecules than hydrophilic amino acids. The success of this approach in modeling preferential binding indicates that it incorporates the important underlying physics of proteins in mixed solvent systems and that the difficulty in quantitative prediction to date can be surmounted by explicitly incorporating the complex protein−solvent and solvent−solvent interactions.

Introduction

Proteins are seldom solvated by pure water. Other solvent components, such as buffer salts and stabilizers, are ubiquitous in the laboratory and in formulations of therapeutic proteins. Similarly, intracellular solutions are crowded with many types of proteins, metabolites, nucleic acids, osmolytes, and other molecules. The presence of these other components, hereafter called "cosolvents", generally alters protein equilibria and reaction kinetics by perturbing the chemical potential of the protein system. Cosolvents perturb the chemical potential of the protein system by associating either more strongly or more weakly with the protein than water. This phenomenon, called "preferential binding",[1] is of great interest because it governs the physical and chemical properties of proteins.

When a cosolvent (X) is added to an aqueous protein solution, it alters the chemical potential of the protein (μ_P) via the following relationship[2]

$$\Delta\mu_P^{tr} = \int_0^{m_X} \left(\frac{\partial \mu_P}{\partial m_X}\right)_{m_P} dm_X \quad (1)$$

$$= -\int_0^{m_X} \left(\frac{\partial \mu_X}{\partial m_X}\right)_{m_P} \left(\frac{\partial m_X}{\partial m_P}\right)_{\mu_X} dm_X \quad (2)$$

where $\Delta\mu_P^{tr}$ is the transfer free energy of the protein from pure water into the mixed solvent system, m is molality, and subscripts X and P identify the cosolvent and protein, respectively. Two partial derivatives appear in eq 2. The first captures the dependence of the cosolvent chemical potential on cosolvent molality and can be evaluated by experiments on a binary mixture of cosolvent and water ($m_P \rightarrow 0$). The second partial derivative is the "preferential binding coefficient", Γ_{XP}

$$\Gamma_{XP} \equiv \left(\frac{\partial m_X}{\partial m_P}\right)_{\mu_X} \quad (3)$$

The preferential binding coefficient is a way in which binding can be defined thermodynamically. It is also particularly useful when binding is weak. The preferential binding coefficient is a measure of the excess number of cosolvent molecules in the domain of the protein per protein molecule (Figure 1). The connection between the thermodynamic definition (eq 3) and the intuitive notion of binding (local excess number of molecules) comes from statistical mechanics, where it can be shown that[3,4]

$$\Gamma_{XP} = \left\langle n_X^{II} - n_W^{II}\left(\frac{n_X^I}{n_W^I}\right)\right\rangle \quad (4)$$

In the above equation, n denotes the number of a specific type of molecule (subscript X for the cosolvent and subscript W for water) in a certain domain (superscript I for a bulk volume outside of the vicinity of the protein and superscript II for a volume in the protein vicinity), and angle brackets denote an ensemble average. Note that Γ_{XP} is independent of the choice of the boundary between the domains, as long as the boundary is far enough from the protein.

If the cosolvent concentration is higher in the vicinity of the protein than in the bulk, Γ_{XP} is greater than zero, and μ_P is lower in the presence of the cosolvent than in its absence. Denaturants such as urea and guanidinium chloride exhibit this type of binding behavior. The reverse is true for sugars, such as trehalose. In trehalose solutions, there is generally a deficiency

* Author to whom correspondence should be addressed. E-mail: trout@mit.edu.

10.1021/jp0363996 CCC: $25.00
Published on Web 11/20/2003

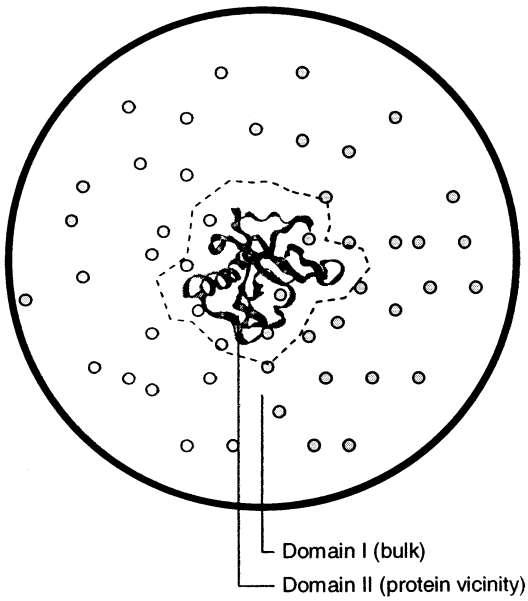

Figure 1. Physical interpretation of the preferential binding coefficient. Interactions of solvent molecules with the protein at the protein−solvent interface generally induce solvent concentration differences in the local (II) and bulk (I) domains. Γ_{XP} is the thermodynamic measure of the number of cosolvent molecules bound to the protein, or in other words, the excess number of cosolvent molecules in the vicinity of the protein vs the number of cosolvent molecules in an equivalent volume of bulk solution.

of trehalose and an excess of water in the vicinity of the protein. For this "preferential hydration" case, Γ_{XP} is less than zero and μ_P is higher in the presence of the cosolvent.

Thirty years ago, Timasheff pioneered the use of high-precision densitometry to measure preferential binding coefficients for protein−cosolvent systems.[2,5−7] More recently, differential scanning calorimetry (DSC)[8] and vapor pressure osmometry (VPO)[9] have been used to the same end. Preferential binding coefficients are rigorous thermodynamic quantities and are related to virial coefficients, activity coefficients, and free energies via standard thermodynamic relations for multicomponent solutions.[10]

Experimental studies by the above methods have led to some generalizations about preferential binding coefficients:

1. Γ_{XP} may be positive or negative, indicating that interactions of the protein and cosolvent are favorable or unfavorable, respectively.

2. Γ_{XP} is proportional to cosolvent molality at low concentration of cosolvent (often as high as $m_X \sim 1$ m and higher).[9,11,12]

3. Γ_{XP} is roughly proportional to the protein−solvent interfacial area.[2]

The second generalization above, together with the fact that many binary mixtures of cosolvent and water ($m_P \rightarrow 0$) are nearly ideal at low concentration of cosolvent, leads to a useful simplification of equation 2

$$\Delta\mu_P^{tr} = -\int_0^{m_X} \left(\frac{\partial RT \ln m_X}{\partial m_X}\right)_{m_P} \left(\frac{\Gamma_{XP}}{m_X}\right) m_X \, dm_X \qquad (5)$$

$$= -RT\left(\frac{\Gamma_{XP}}{m_X}\right) \int_0^{m_X} dm_X \qquad (6)$$

$$= -RT\Gamma_{XP} \qquad (7)$$

Equation 7 provides a simple and convenient link between preferential binding coefficients and free energies. This relation leads to the useful rule that when Γ_{XP} is proportional to m_X, for each cosolvent molecule that preferentially interacts with the protein the protein's free energy is reduced by approximately 0.6 kcal/mol at 25 °C. The simplicity of this relation is a natural result of the close relationship between Γ_{XP} and a second virial coefficient.

To be able to predict preferential binding coefficients and understand their origins, the above thermodynamic framework and general observations must be augmented by a mechanistic model. Several such models have been presented in the literature, including models based on the binding polynomial or statistical mechanical partition function, solvent−cosolvent exchange at defined sites, cosolvent partitioning between the local and bulk domains, and group contribution methods for estimating transfer free energies.

The most general model of cosolvent binding hitherto presented comes from considering an equilibrium of all possible protein−cosolvent complexes, from which it can be shown that[13]

$$\Delta\mu_P^{tr} = -RT \ln(1 + \sum_i \sum_j K_{ij} m_W^i m_X^j) \qquad (8)$$

where K_{ij} is the equilibrium constant for a reaction of a protein molecule, i molecules of water, and j molecules of cosolvent into a complex. While this model is completely general, its utility is limited because it is not possible to determine experimentally the many K_{ij} parameters present in eq 8.

Schellman's *site exchange model*[4] provides a way to simplify this general expression to a form containing a single parameter. This model treats binding as a family of protein−solvent exchange reactions such as

$$P{\cdot}W_i + X \rightarrow P{\cdot}X + iW \qquad (9)$$

where P is the protein, W is water, X is cosolvent, and i is the exchange stoichiometry. The simplification requires the assumptions that 1:1 exchange reactions ($i = 1$) occur on a fixed number of identical, independent sites and that the sites are far from saturation with cosolvent (i.e., the apparent dissociation equilibrium constant for each site is well above the cosolvent concentration). The number of sites, n, is approximated by the number of water molecules present in a monolayer around the protein. These simplifications reduce eq 8 to

$$\Delta\mu_P^{tr} = -nRT\langle K\rangle m_X \qquad (10)$$

where $\langle K\rangle$ is the average equilibrium constant of binding at a single site. The single parameter $\langle K\rangle$ can then be determined from an experimental measurement of Γ_{XP}. When eq 7 holds, the relation between $\langle K\rangle$ and Γ_{XP} is simply

$$\langle K\rangle = \Gamma_{XP}/nm_X \qquad (11)$$

Values of $\langle K\rangle$ for different proteins in this linear regime are roughly equal.[14] $\langle K\rangle$ cannot, however, be determined without knowledge of Γ_{XP} or other free-energy data on the particular cosolvent system of interest. In fact, one can say that $\langle K\rangle$ is defined by Γ_{XP}.

Another model that recasts preferential binding coefficient data in terms of a single model parameter is the *local-bulk domain model* developed by Courtenay et al.[9] The parameter in this model is the partition coefficient K_P, relating the number of water molecules and cosolvent molecules in the local and bulk domains via

$$K_P = \frac{n_X^{II}/n_W^{II}}{n_X^{I}/n_W^{I}} \tag{12}$$

Similar to the site exchange model, the convention used in this model is that the local domain consists of a monolayer of water and enough cosolvent to obtain the experimentally observed Γ_{XP}. Note that because the absolute occupancy of water and cosolvent in the local domain cannot be easily determined by experiment, the local-bulk domain model effectively defines n_W^{II}. Like $\langle K \rangle$, values of K_P can be used to predict Γ_{XP} at other cosolvent concentrations or for other proteins in the same cosolvent, but predictions cannot be made in the absence of Γ_{XP} or free-energy data on the same cosolvent system.

Last, *transfer free-energy models*, pioneered by Bolen's group,[15] take a different approach. These models conceptually divide whole proteins into groups[16] such as the amino acid side chains and the protein backbone and model the transfer free energy of the whole protein as a sum of the transfer free energy of the groups it comprises, via

$$\Delta\mu_P^{tr} = \sum_i \alpha_i \Delta g_i^{tr} \tag{13}$$

where Δg_i^{tr} is the transfer free energy of the model group and α_i is the solvent accessible area of the group in the whole protein, normalized to the solvent accessible area of the model compound. The overall $\Delta\mu_P^{tr}$ can then be predicted for any system of known structure. In the context of the previously described models, the transfer free-energy model can be thought of as a linearized binding model where each surface group or amino acid in the protein represents a different type of independent binding site, and the binding constants for those sites are determined by experiments on model compounds, such as free amino acids or cyclic diamino acid compounds. Predictions made by transfer free-energy models have met with mixed success. A linear group contribution model (eq 13) may be too simple to capture all of the important contributions to $\Delta\mu_P^{tr}$.[17]

While the above models have helped in the understanding of the phenomenon of preferential binding, they generally incorporate strong assumptions, and they necessitate the use of experimental data on highly analogous systems in order to determine model parameters and make predictions. Thus, their uses as predictive tools and as tools to gain insight into specific systems are limited.

In this work, we developed a predictive, molecular-level approach for the study of preferential binding based on all-atom, statistical mechanical models that use no adjustable parameters. To date, statistical mechanical models of preferential binding have only been developed for interactions of ions with charged cylinders[18,19] and for interactions of two-dimensional "hard circles" with a linear interface,[20] both far too simple to be generally applied to protein−cosolvent systems. Other explicit mixed solvent simulations of proteins and amino acids have been performed,[21−25] but these studies did not compute thermodynamic quantities related to preferential binding. In our approach, we define the number of "bound" molecules in a thermodynamically consistent way and do not a priori incorporate any information about "binding sites". The use of our approach for the computation of preferential binding coefficients was validated in two systems by comparison with experimental data from the literature. Additionally, the molecular-level detail of the approach provides new insights into the following issues:

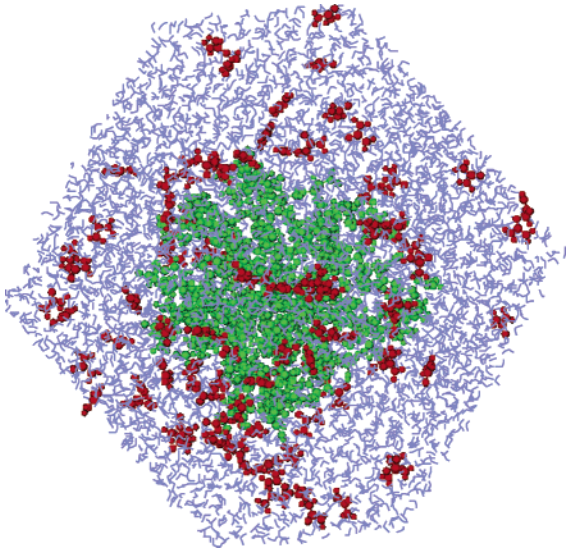

Figure 2. A simulation cell containing RNase T1 (center, green spheres) solvated by water (thin blue lines) and urea (red spheres). Figure generated with VMD.[28]

1. The changes in solvent and cosolvent concentration as a function of distance from the protein surface.
2. A precise definition of the "local domain" (Figure 1).
3. The differences in preferential binding or apparent binding equilibrium constant at different locations on the protein−solvent interface.

The success of this method in modeling preferential binding indicates that it captures the important underlying physics of protein−cosolvent−water systems and that the difficulty in quantitative prediction to date can be surmounted by explicitly incorporating the complex protein−solvent and solvent−solvent interactions.

A New, Molecular-Level Approach to Computing Preferential Binding. Our approach uses explicit atomic interaction potentials (force fields), such as Lennard-Jones, Coulombic, spring, and torsion interactions, with prefit coefficients.[26,27] Thermodynamic properties, such as preferential binding coefficients, are computed by averaging in the time domain via molecular dynamics (MD). A snapshot from a dynamic simulation of RNase T1 in a urea solution is shown in Figure 2. The results of such simulations contain all of the information needed to extract thermodynamic properties such as Γ_{XP}.

Molecular dynamics uses Newton's second law of motion, that acceleration is the quotient of force and mass, to compute the positions of each atom in the system as a function of time. To do this, an energy model, sometimes called a "force field", that can be used to compute the net force on any atom in any configuration is employed.

During the MD run, the positions of each atom are recorded at fixed intervals in time. These "snapshots" form an ensemble of configurations which can then be used to compute thermodynamic properties, such as Γ_{XP}.

Importantly, this method of computing Γ_{XP} does not introduce any adjustable parameters to model preferential binding or any other aspect of a system containing a protein and two solvent components. All of parameters required by the MD method for energy computations are determined independently of this particular modeling objective and in fact have been shown to be generally applicable to biological systems.[29] Thus, the method developed here could be used to estimate Γ_{XP} and $\Delta\mu_P^{tr}$ in systems where no experimental data is available. It therefore facilitates the study of preferential binding when direct experi-

mental study is difficult, such as at transition state configurations or at marginally stable states of proteins. Furthermore, it yields detailed, local, molecular-level insight into the system studied.

Another benefit of this approach is that when eq 7 holds (such as for urea and glycerol), the protein transfer free energy ($\Delta\mu_P^{tr}$) can be calculated from a single Γ_{XP} simulation. Traditional free-energy calculation methods such as thermodynamic integration[30,31] require 15−20 trajectories, which is computationally difficult for protein systems of this size.

Preferential Binding Coefficients of Constituent Groups. Because proteins have a range of different functional groups in different orientations on their surfaces, the concentrations of solvents and cosolvents near different patches on the protein's surface may be different. For example, the vicinity of a hydrophobic patch on the protein may have a lower concentration of water and a higher concentration of cosolvent than in the vicinity of a hydrophilic patch. Preferential binding experiments capture only the average effect arising from all of the interactions over the entire protein−solvent interface; however, molecular simulations allow more detailed analyses of the local contributions to preferential binding coefficients.

A protein can be thought of as a set of nonoverlapping constituent groups,[16] each of which has its own preferential binding coefficient defined by the composition of the solvent in its immediate vicinity. Similar to group contribution methods for computing transfer free energies (see Introduction), one possible group definition is that each type of amino acid side chain (up to 20) and the amino acid backbone are distinct groups. To compute a preferential binding coefficient for a constituent group, the solvent molecules in the local domain are assigned only to the nearest group (i) and the "group preferential binding coefficients" ($\Gamma_{XP,i}$) can be defined as

$$\Gamma_{XP,i} \equiv \left\langle n_{X,i}^{II} - n_{W,i}^{II}\left(\frac{n_X^I}{n_W^I}\right)\right\rangle \tag{14}$$

where $n_{X,i}^{II}$ and $n_{W,i}^{II}$ are the number of cosolvent and water molecules in the local domain that are nearest to group i. If each solvent molecule in the local domain is assigned to a group, the overall preferential binding coefficient is simply the sum of all of the group preferential binding coefficients

$$\Gamma_{XP} = \Sigma\Gamma_{XP,i} \tag{15}$$

The group preferential binding coefficients decompose the effect of each small subset of the protein on the overall preferential binding coefficient. This is analogous to the group contribution models for transfer free energy except that the parameters are extracted from a simulation of an entire protein instead of experiments on model compounds.

Minimum Simulation Time. Sufficient sampling of position−space configurations in time is required for the accurate calculation of Γ_{XP} via eq 3. Assuming that the average protein solution structure is close to that of the initial (crystal) structure and that water molecules sample position space rapidly because of their high density, the most important time scale to be captured is that of the cosolvents sampling position space. One way to estimate this time is that it must be much larger than the average time between cosolvent−cosolvent contacts.

An estimate of the time between contacts can be obtained as

$$t_{contact} \approx \frac{1}{12\mathscr{D}}\left(\frac{V_{solv}}{n_X}\right)^{2/3} \tag{16}$$

where $\mathscr{D}$ is the cosolvent diffusivity, V_{solv} is the solvent volume, and n_X is the number of cosolvent molecules. For the simulations performed here, the solvent is mostly water, so eq 16 can be further simplified to yield

$$t_{contact} \approx \frac{1}{12\mathscr{D}}\left(\frac{1}{N_A\rho_W m_X}\right)^{2/3} \tag{17}$$

where N_A is Avogadro's number and ρ_W is the density of water in kg/m^3. For a 1 m cosolvent in water system with a cosolvent diffusivity of 2×10^{-9} m^2/s (a lower bound on the diffusivities of the cosolvents studied here), $t_{contact}$ is about 30 ps. Thus, nanosecond trajectories will be required for good sampling of cosolvent position space. Importantly, this time increases as the cosolvent concentration decreases, implying that there is a minimum concentration that can be studied with any given amount of computational resources.

TABLE 1: Details of Four MD Simulations Performed[a]

cosolvent	protein	T (°C)	pH	n_X	n_W	$\langle l\rangle$ (Å)
urea	RNase T1	25	7	90	4274	57.48
glycerol	RNase T1	25	7	87	4582	59.24
glycerol	RNase A	25	3	90	5480	62.86

[a] n_X is the number of cosolvent molecules; n_W is the number of water molecules; and $\langle l\rangle$ is the average dimension of the primary unit cell (which varies during the run at constant pressure).

Methodology

Molecular Simulations. Molecular dynamics was used to sample the phase space of proteins solvated by water and a cosolvent. Version 28 of the CHARMM[26] molecular dynamics package was used for all simulations. The CHARMM force field was used for the protein, and the TIP3P model[32] was used for water. A force field was constructed for glycerol using the standard CHARMM geometries and partial charges for the atoms in a −CHOH− unit.[26,27] Urea was assumed to be planar with bond lengths equal to the CHARMM standards and partial charges recomputed as done previously[33] while using the CHARMM van der Waals mixing rules in the objective function.

The structures of RNase A (PDB code: 1fs3) and RNase T1 (PDB code: 1ygw) were obtained from the Protein Data Bank.[34] In total, three simulations were performed: RNase A in 1 m glycerol (pH 3), RNase T1 in 1 m glycerol (pH 7), and RNase T1 in 1 m urea (pH 7). Details of each simulation are shown in Table 1. Each protein was solvated in a trucated octahedral box extending a minimum of 9 Å from the protein. The pH of each simulation was fixed by setting the protonation states of each ionizable side chain to the dominant form expected for each amino acid at the pH of interest. Arginine, cysteine, lysine, and tyrosine were protonated in all of the simulations. Aspartate, glutamate, and histidine were assumed to have pK_a values of 3.4, 4.1, and 6.6,[35,36] respectively, and were therefore protonated in the simulation at pH 3 and deprotonated at pH 7. Initial placement of water and cosolvent molecules were random. Protein counterions were placed using SOLVATE 1.0. The system was first energy minimized at 0 K, next heated to 298.15 K, and then equilibrated for 1 ns in the NTP ensemble at 1 atm. For the computation of the properties of interest, 2 ns of dynamics were then run, during which statistics were computed from snapshots of the trajectory every picosecond.

Calculation of Preferential Binding Coefficients. The trajectories were then used to define the local and bulk regions and compute Γ_{XP} in the following manner. For the purpose of computing Γ_{XP} and other thermodynamic and structural parameters, each water and cosolvent molecule was treated as a point

at its center of mass. The distance of each of these points to the protein's van der Waals surface was computed, and then $\rho_W(r)$ and $\rho_X(r)$, defined as the number densities of these points at a distance r from the protein, were computed. In all cases, the $\rho(r)$ functions exhibited peaks and valleys characteristic of solvation shells in the range $0 < r < 6$ Å. At distances in the range of 6−8 Å and higher, such variations are no longer seen and the local number density is defined as bulk number density, $\rho(\infty)$. Such a region far from the protein containing a spatially uniform concentration of water and cosolvent must be present in the simulation cell in order to define the local and bulk regions and calculate Γ_{XP}.

The position of the boundary between the local and bulk domains, a distance of r^* away from the surface of the protein, was then determined by choosing the minimum distance at which no significant difference between $\rho(r^*)$ and $\rho(\infty)$ was apparent for either water or cosolvent. All solvent molecules whose centers of mass fell inside a distance of r^* from the protein's van der Waals surface were defined as belonging to the local domain (II), and all other solvent molecules were defined as belonging to the bulk domain (I). With these definitions of the domains, the instantaneous preferential binding coefficient, $\Gamma_{XP}(t)$, was computed as

$$\Gamma_{XP}(t) \equiv n_X^{II} - n_X^{I}\left(\frac{n_W^{II}}{n_W^{I}}\right) \tag{18}$$

for each time point in each trajectory. The preferential binding coefficient, Γ_{XP}, was then computed for each trajectory as the time average of these instantaneous values

$$\Gamma_{XP} = \frac{1}{t}\int_0^t \Gamma_{XP}(t')\,dt' \tag{19}$$

The radial distribution functions $g_X(r)$ and $g_W(r)$ are defined as

$$g_i(r) \equiv \rho_i(r)/\rho_i(\infty) \tag{20}$$

where i represents water (W) or a cosolvent (X) species. These functions provide another route to compute Γ_{XP}

$$\Gamma_{XP} = \langle n_X^{II}\rangle - \left\langle\left(\frac{n_X^{I}}{n_W^{I}}\right)n_W^{II}\right\rangle \tag{21}$$

$$= \rho_X(\infty)\int g_X\,dV - \left(\frac{\rho_X(\infty)}{\rho_W(\infty)}\right)\rho_W(\infty)\int g_W\,dV \tag{22}$$

$$= \rho_X(\infty)\int (g_X - g_W)\,dV \tag{23}$$

where each integral is over the local domain or the entire system (since $g_X - g_W = 0$ in the bulk domain).

The boundary between domains I and II must be placed far enough from the protein to ensure that it is in the bulk, yet at the smallest such distance so that statistical fluctuations in the number of molecules in the domains can be minimized. We can use the values of $g_X(r)$ and $g_W(r)$ to determine the optimal boundary. Defining Γ_{XP}^* as the apparent preferential binding coefficient resulting from defining the local domain as those molecules whose centers of mass lie inside a distance r^* from the protein

$$\Gamma_{XP}^*(r^*) = \rho_X(\infty)\int_0^{r^*}(g_X - g_W)\frac{dV}{dr}\,dr \tag{24}$$

The error in Γ_{XP}, E_Γ, introduced by selecting a particular value of r^* is then

$$E_\Gamma = \Gamma_{XP}^*(r^*) - \Gamma_{XP} \tag{25}$$

$$= -\rho_X(\infty)\int_{r^*}^{\infty}(g_X - g_W)\frac{dV}{dr}\,dr \tag{26}$$

When r^* is selected properly, the surface defined by $r = r^*$ is entirely in the bulk solution, $g_X(r^*) = g_W(r^*) = 1$, and $E_\Gamma = 0$. Thus, selecting r^* as the minimum distance for which all $r \geq r^*$ satisfy $g_X(r) = g_W(r) = 1$ (within the error of the simulation) is optimal.

Calculation of Constituent Group Preferential Binding Coefficients. For each simulation, up to 21 constituent group preferential binding coefficients were calculated. The 21 groups were each type of amino acid side chain present in the protein (up to 20) and the protein backbone. The "protein backbone" was defined as the −NH−CH−COO− unit, as well as the two extra protons at the N-terminus and extra oxygen atom at the C-terminus of the protein. The glycine side chain was defined as the proton bound to the alpha carbon that would be replaced by a substituent to form a different L-amino acid.

For the simulation of RNase T1 in glycerol solution, the constituent group preferential binding coefficients for the 15 individual serine residues in the protein were also calculated. For this calculation, solvent and cosolvent molecules that were nearest to an atom in the protein that was not part of a serine side chain were not considered.

Water and cosolvent molecules were associated with a specific constituent group by computing the distance from the center of mass of each solvent molecule to the van der Waals surface of every atom in the protein, selecting the protein atom that was nearest to the solvent molecule, and then determining to what constituent group this nearest protein atom belonged.

Estimation of Statistical Error. The statistical error arising from computing averaged properties from a finite trajectory was estimated in the following fashion:

1. The dynamic trajectory of interest was divided into n pieces.
2. The mean of the property of interest was computed in each piece. These means were designated $\bar{x}_i$ where $i = 1, ..., n$.
3. The standard deviation of the $\bar{x}_i$ values was computed.
4. This standard deviation was divided by $n^{1/2}$, and the quotient was designated σ_m, an estimate of the error in the mean determined by time averaging the full trajectory.

The number of pieces n into which the trajectory is divided must be small enough to ensure that the means of each piece (the $\bar{x}_i$) are statistically independent. An autocorrelation analysis (not shown) of several trajectories of $\Gamma_{XP}(t)$ data and the underlying molecular counts (n_i^I and n_i^{II}) indicates that a window of about 0.2 ns is sufficiently large for this to be true. Therefore, for a 2 ns dynamics trajectory, a value of $n = 2/0.2 = 10$ was used.

For long trajectories, the statistical error σ_m is roughly proportional to the inverse square root of the trajectory length. This property can be used to estimate the trajectory length required to achieve a given level of statistical accuracy after a small trajectory has been generated and analyzed.

Results and Discussion

Radial Distribution Functions of Water and Cosolvents. The radial distribution functions of water, urea, and glycerol were computed for all three simulations as described in Methodology and are shown in Figure 3.

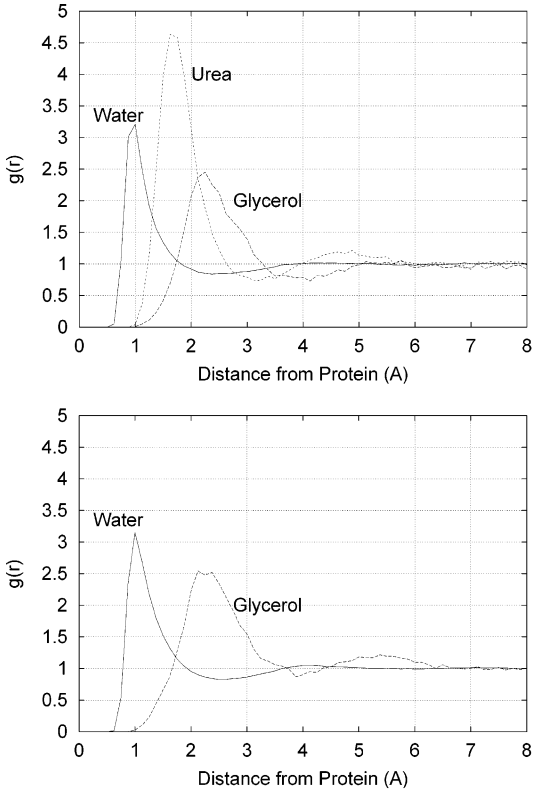

Figure 3. Radial distribution functions of water, urea, and glycerol are shown for simulations of RNase T1 in glycerol and urea solutions (left) and RNase A in a glycerol solution (right). In the left-hand figure, the difference between the two $g_W(r)$ functions is not visible at this scale.

At very short distances, $r < 0.6$ Å for water and $r < 1.0$ Å for glycerol and urea, regions of total solvent and cosolvent exclusion due to very strong van der Waals repulsion can be seen. The size of these "totally excluded" regions is much smaller than one would expect based on the apparent van der Waals radii of the solvent and cosolvent molecules alone (for example, $r \approx 1.5$ Å for water and 2.2 Å for urea),[37] indicating that electrostatic attractive forces play an important role in solvation even at these distances. After the regions of total exclusion, strong first-coordination shells of these three molecules can be clearly seen. The peaks of the first-coordination shells become more distant from the protein as the size of the molecules they correspond to increases. Significantly smaller second-coordination shell peaks are also visible for urea solvating RNase T1 and glycerol solvating RNase A. At distances greater than 6−7 Å from the protein, solvation shells cannot be discerned, and the number densities of water, urea, and glycerol reach their bulk values.

In the simulations of RNase T1 in glycerol and urea solutions, the radial distribution functions for glycerol and urea are quite different. The maximum value of $g_X(r)$ for urea is over 4.5 while that for glycerol is about 2.5. The difference in these maximum values, while significant, is not sufficient to say that the number of urea molecules coordinated to the protein (n_X^{II}) is higher than the number of glycerol molecules coordinated; this can only be done by integrating each $g_X(r)$ function appropriately via eq 23.

The radial distribution functions for both water and glycerol are similar in the simulations of RNase A and RNase T1 in glycerol solution despite the fact that the proteins and the pHs of the solutions are different. Given that the proteins are of similar size, this observation is consistent with the fact that the values of Γ_{XP} for the two solutions are close.

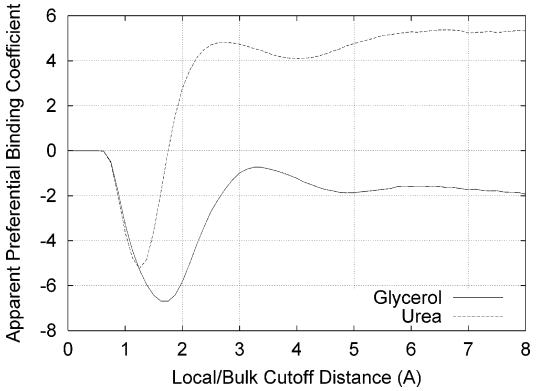

Figure 4. Apparent preferential binding coefficient as a function of the cutoff distance between the local and bulk domains for simulations of RNase T1 in glycerol and urea solution.

TABLE 2: Preferential Binding Coefficients Computed from MD Simulations and Compared with Available Experimental Data at Similar Cosolvent Concentrations[a]

system	m_{bulk}	simulation Γ_{XP}	experiment Γ_{XP}
urea/RNase T1	1.10 m	5.2 ± 1.0	6.4[39]
glycerol/RNase T1	1.07 m	-1.6 ± 0.8	
glycerol/RNase A	0.91 m	-0.9 ± 1.0	-1.7 ± 0.8[6]

[a] A wide range of behavior (positive and negative preferential binding coefficients) can be modeled without the use of adjustable parameters. The confidence intervals on Γ_{XP} (MD) are an estimate of the statistical error resulting from the use of a finite trajectory. For easier comparison, the experimental values of Γ_{XP} reported above were interpolated to m_{bulk} from data sets spanning the molality of interest

Preferential Binding Coefficients. The radial distribution functions in Figure 3 suggest that r^* in the range of 6−8 Å is an appropriate choice of boundary between the local and bulk domains. The error in Γ_{XP} introduced by a particular choice of the boundary distance, r^*, can be estimated by plotting the apparent preferential binding coefficient (Γ_{XP}^*) vs r^* (Figure 4). Γ_{XP}^* depends very strongly on r^* in the first solvation shell ($r = 0-4$ Å) and weakly on r^* in the second solvation shell ($r = 4-6$ Å). In the range $r = 6-8$ Å, the dependence of Γ_{XP}^* on r^* is small (±0.5) and is less than the statistical error in Γ_{XP} (shown in Table 2, explained below). Therefore, a cutoff distance of 6 Å, or about two solvation shells, is sufficiently large to minimize systematic error in Γ_{XP} caused by the choice of r^*. If only a single solvation shell were considered ($r^* \sim 3.5-4$ Å), a systematic error in Γ_{XP} of approximately 0.5−1 molecules would be introduced as a result of neglect of the second solvation shell.

The preferential binding coefficient, Γ_{XP}, was computed via eq 3 using $r^* = 6$ Å as the boundary between the local and bulk domains. A confidence interval for this ensemble average was computed as described in Methodology. The binding coefficients and their statistical uncertainties are shown in Table 2. Experimental values from the literature were available for two out of three of these protein−cosolvent systems, and our computed values of Γ_{XP} agree quite favorably with these. The fact that this occurs for both positive and negative values of Γ_{XP} without the use of any adjustable parameters is very encouraging. For a cosolvent that obeys eq 7, the confidence intervals of ±1.0 in Γ_{XP} represent a confidence limit in the transfer free energy of about 0.6 kcal/mol, which is a typical value for free energies calculated via this type of molecular

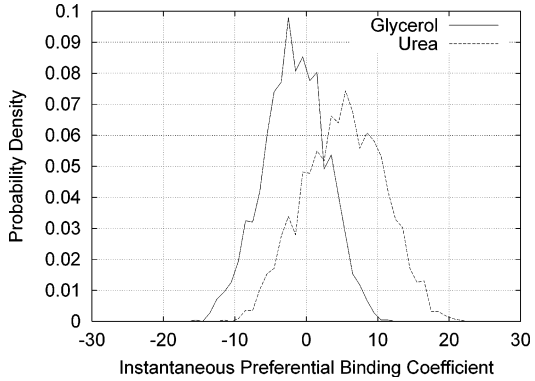

Figure 5. $\Gamma_{XP}(t)$ probability density function. A wide range of values of $\Gamma_{XP}(t)$ are sampled as water and cosolvent molecules diffuse between the local and bulk domains.

simulation. Achievement of this level of accuracy despite the fact that structural fluctuations in the native state ensemble of proteins have been observed on much longer time scales[38] than the time scale of the simulations performed here suggests that solvent dynamics are more important than protein structural dynamics in determining Γ_{XP}.

$\Gamma_{XP}(t)$ probability density functions for the simulations of RNase T1 in urea and glycerol solution are shown in Figure 5. The range of instantaneous values of the preferential binding coefficient, $\Gamma_{XP}(t)$, is quite large relative to the absolute values of Γ_{XP}. $\Gamma_{XP}(t)$ values in excess of $\Gamma_{XP} \pm 15$ are observed. The breadths of these distributions are related to the size of the interface between the local and bulk domains and indicate the importance of sampling a large number of solvent configurations to obtain the macroscopic, averaged Γ_{XP} (eq 19).

The Relation between Solvent Accessible Area and the Number of Molecules in the Local Domain. The solvent-accessible areas of whole proteins (SAA) and constituent groups (SAA_i) in crystal structures have been used extensively in analyzing proteins. SAA and SAA_i are essentially simple ways of measuring water coordination numbers. In models developed to date, SAA or SAA_i has been used to estimate n_W^{II} or $n_{W,i}^{II}$ by assuming that the local domain is a monolayer of water and each water molecule occupies approximately 10 Å^2 of the solvent-accessible area. Since we have introduced a new notion of the local domain, it is worthwhile to see what relationships exist between SAA_i and the coordination numbers $n_{W,i}^{II}$ and $n_{X,i}^{II}$ that utilize this definition.

A scatter plot of the solvent-accessible area of a set of constituent groups (amino acid side chains and the protein backbone) vs the number of water molecules in the local domain for three different simulations is shown in Figure 6. Solvent-accessible area was calculated analytically in CHARMM (based on Richmond's method)[40] using a 1.4-Å probe. There is a strong, linear correlation of these variables with a slope of 4.2 Å^2/molecule and correlation coefficient of 0.96. Similarly strong correlations are seen for SAA_i with $n_{X,i}^{II}$ in individual simulations. A summary of proportionality constants and correlation coefficients for these relationships is shown in Table 3. If the time-average SAA_i from each dynamics simulation is used instead of the crystal-structure SAA_i values, the correlation coefficients increase slightly. Because the time-average solvent-accessible areas are higher than those in the crystal structure, the proportionality constants shown in Table 3 also increase.

Constituent Group Preferential Binding Coefficients. The constituent group preferential binding coefficients were calculated for each simulation as described in Methodology and are shown in Figures 7−10 as the number of water and cosolvent

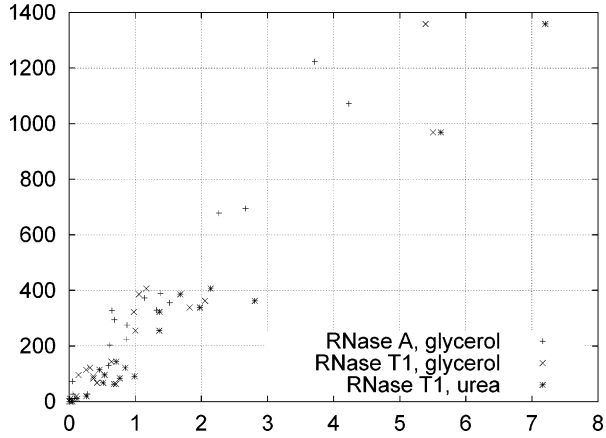

Figure 6. Correlation of solvent-accessible area and the number of water molecules in the local domain of constituent groups. Each point represents a constituent group of either a type of amino acid side chain or the protein backbone in one of the three simulations shown in Table 2. The solvent accessible area of a constituent group and the number of water molecules in the local domain of the solvent near the group ($n_{W,i}^{II}$) are highly correlated.

TABLE 3: Relationships between Solvent-Accessible Area in Each Protein Crystal Structure and Number of Solvent Molecules in the Local Domain for Different Protein−Cosolvent Systems[a]

species (i)	protein	avg protein SAA/n_i^{II} (Å^2/molecule)	r^2
water	RNase A/T1	4.2	0.96
0.91m glycerol	RNase A	290	0.96
1.07m glycerol	RNase T1	230	0.93
1.10m urea	RNase T1	170	0.98

[a] r^2 symbolizes the correlation coefficient.

molecules coordinated to each constituent group. In each figure, a line at the bulk solution composition is also plotted, enabling a quick determination of the composition of the solvent in the vicinity of a constituent group compared to the bulk solvent. The statistical uncertainties in the values of $n_{W,i}^{II}$ and $n_{X,i}^{II}$ (and consequently $\Gamma_{XP,i}$) are high. Because of these uncertainties, we will not report specific values of the group preferential binding coefficients but rather classify them into broad categories based on their statistical likelihood of being either positive, negative, or zero/indeterminate.

The average number of water and glycerol molecules coordinated to each of the 15 serine residues in RNase T1 are shown in Figure 7. A wide range of binding behavior can be seen among the serine residues, all of which have a good degree of solvent exposure. Ser17, 35, and 72 fall above the bulk concentration line and have positive preferential binding coefficients, Ser63 falls below the line and has a negative preferential binding coefficient, and the preferential binding coefficients of the remaining 11 serine residues are not statistically different from zero. The wide range of local concentrations in the vicinities of these serine residues indicates that developing a group contribution method to estimate Γ_{XP} or $\Delta\mu_P^{tr}$ based on primary sequence information and solvent accessibility ($n_{W,i}^{II}$) alone may be difficult. In addition to the type of amino acids present at the protein−solvent interface, other effects such as specific combinations of residues and secondary or tertiary structure must be important in determining water and cosolvent binding behavior. These factors probably contribute to the range of local concentrations seen in Figure 7. For example, Ser35 and Ser72 are proximal to each other and several Gly and Tyr side chains (Gly34, 70, and 71 and Tyr68), which tend to have

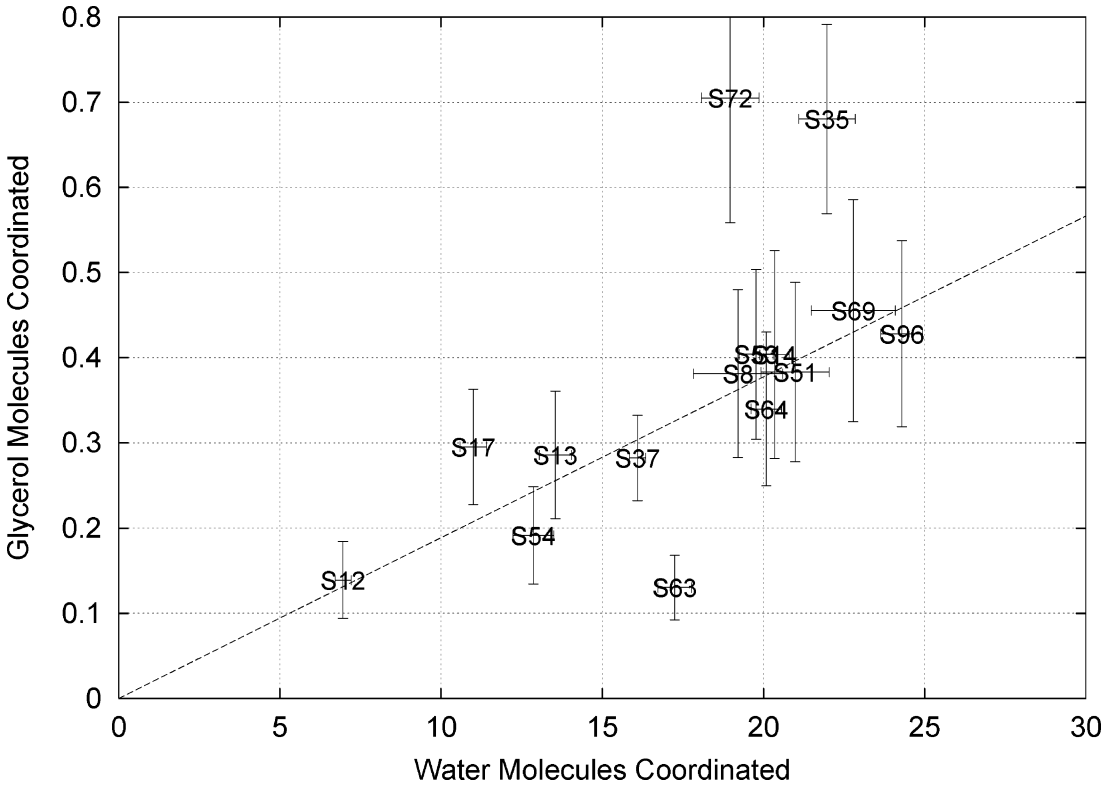

Figure 7. Binding behavior of glycerol and water with the 15 serine residues in RNase T1 is shown as a plot of the number of glycerol molecules in the local domain of each serine residue versus the number of water molecules in the same volume. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line represents the bulk glycerol composition. Ser17, 35, and 72 have positive preferential binding coefficients, Ser63 has a negative preferential binding coefficient, and the remaining 11 serine residues have essentially zero values for their preferential binding coefficients.

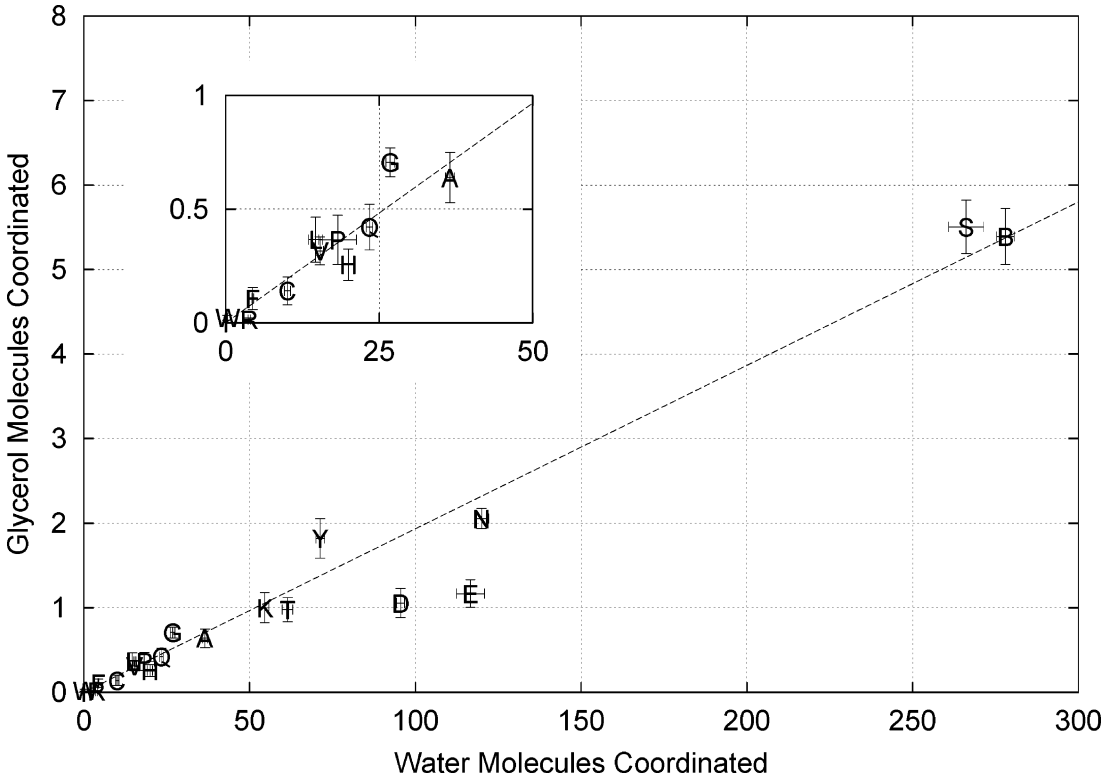

Figure 8. Local binding behavior of urea and water with the amino acid backbone and side chains in RNase T1. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk urea concentration. In addition to the protein backbone and Ser, the hydrophobic amino acids Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water.

positive preferential binding coefficients in glycerol (Figure 9). This may be the reason that the group preferential binding coefficients for these residues are higher than those of the other serine residues.

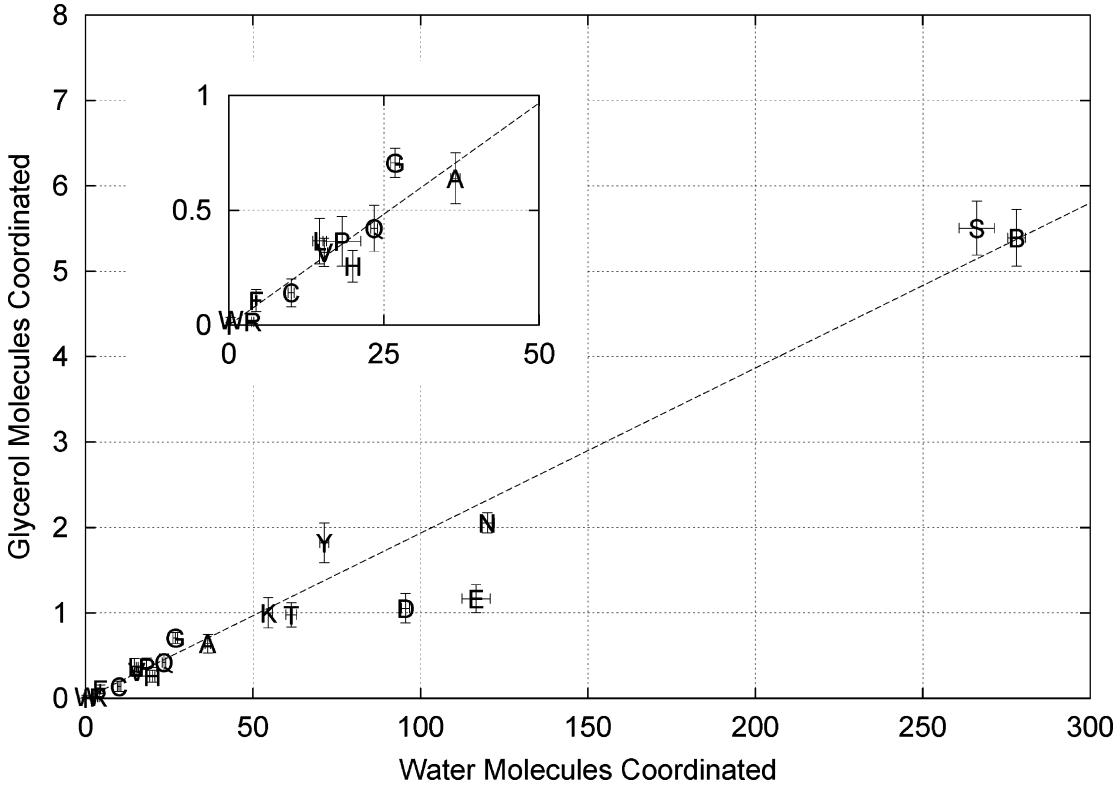

Figure 9. Group preferential binding coefficients for glycerol with the amino acid backbone and side chains in RNase T1. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk glycerol concentration. Tyr and Gly preferentially bind glycerol; Asp and Glu preferentially bind water; and the binding coefficients of the other groups are not statistically different from zero.

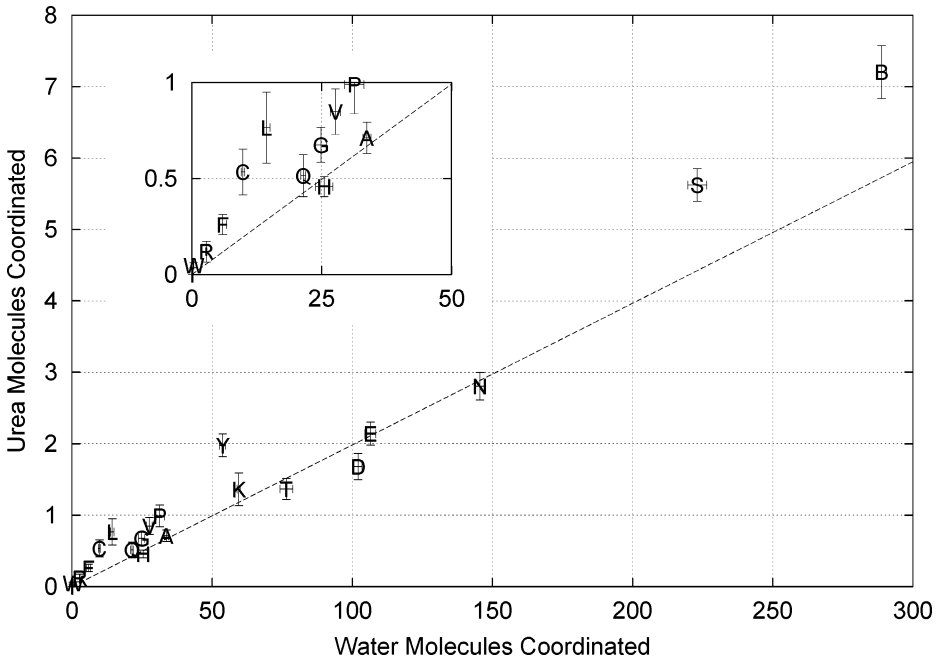

Figure 10. Local binding behavior of glycerol with the amino acid backbone and side chains in RNase A. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk glycerol concentration. All of the constituent groups in RNase A either preferentially bind water or are neutral.

The preferential binding behavior of urea and glycerol with each type of amino acid in RNase T1 and the protein backbone are shown in Figures 8 and 9. In urea solution, the protein backbone and Ser as well as the hydrophobic amino acid side chains of Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water. In glycerol solution, only Tyr and Gly preferentially bind glycerol and Asp and Glu preferentially bind water. Qualitatively, the binding behavior of the amino acid side chains of RNase T1 follow a hydrophobic series, with the hydrophobic side chains tending to bind more cosolvent and the hydrophilic ones tending to bind more water.

The binding behavior of glycerol and water with the amino acid side chains and backbone in RNase A, shown in Figure 10, is significantly different than the binding behavior of these solvent components with the same constituent groups in RNase

T1. (Note that the protonation states of Asp, Glu, and His are different in the two simulations.) The amino acid backbone, which occupies a large fraction of the protein−solvent interface as indicated by its high value of $n_{W,i}^{II}$, has a binding coefficient near zero in RNase T1 and a significant negative binding coefficient in RNase A. More strikingly, Tyr in RNase T1 preferentially binds glycerol whereas Tyr in RNase A preferentially binds water. This is likely because the six Tyr residues in RNase A are at or near the solvent interface (a more hydrophilic region) whereas the nine in RNase T1 are mostly buried (a more hydrophobic region). This difference in solvent exposure is evident from the crystal structures of the proteins but also can be discerned by comparing the water coordination numbers for Tyr in the two proteins: $n_{W,i}^{II}$ for Tyr in RNase A is higher than in RNase T1, even though there are 50% more Tyr residues in RNase T1.

On the basis of the above observations, some generalizations about the effects that these cosolvents have on protein folding equilibria can be postulated, the validity of which must be confirmed via future studies. In urea solution, most of the constituent groups in RNase T1 either preferentially bind urea or are indifferent to urea and water. Asp, which is found on the surface of RNase T1, is the only constituent group that is significantly below the bulk concentration line in Figure 8 and therefore preferentially binds water over urea. Since the amino acids that compose the core of RNase T1 and are exposed upon unfolding preferentially bind urea, this pattern suggests that the preferential binding coefficient or urea with unfolded RNase T1 is higher than that with native RNase T1. This is thermodynamically consistent with urea's well-known ability as a denaturant. Inversely, in glycerol solution, almost all of the constituent groups in RNase A and T1 are neutral or preferentially bind water. This is consistent with the fact that glycerol binds less to the unfolded protein than the native state, and therefore is a protein stabilizer. Both of these generalizations are consistent with earlier work on model compounds.[17]

Conclusions

A quantitative method based on molecular dynamics simulations using all atom potential models has been developed and validated for calculating preferential binding coefficients. Our method is not a derivative of thermodynamic integration or thermodynamic perturbation methods and requires only a single trajectory to compute the transfer free energy of a protein into a weak-binding cosolvent system. Our results match experimental data well for glycerol and urea solutions, covering a range of positive and negative binding behavior. This work also augments experimentally observable, macroscopic thermodynamics with the mechanistic insight provided by a molecular-level, statistical mechanical model.

Variations in the radial distribution functions with distance for each cosolvent are evident up to about 6 Å, or two solvation shells of water, away from the protein. Glycerol is not totally excluded from close contact with the protein, but glycerol is less likely than urea to be found in such a position. The radial distribution functions of water and cosolvents are sufficient to calculate preferential binding coefficients by integrating over a suitable solvent volume.

The binding behavior of the amino acid side chains in RNase T1 qualitatively follow a hydrophilic series, with more hydrophilic amino acids in the protein tending to have a higher concentration of water in their vicinity. The constituent group binding behavior differs between the groups in RNase A those in RNase T1. Development of a group contribution method at the amino acid level for estimating binding coefficients or transfer free energies of whole proteins is complicated by the wide range of coordination behaviors observed for single types of amino acids in different environments on the protein surface.

Acknowledgment. The authors wish to acknowledge Prof. D. I. C. Wang for very helpful discussions and the National Institutes of Health Biotechnology Training Program and the National University of Singapore for funding.

References and Notes

(1) Timasheff, S. N. *Adv. Protein Chem.* **1998**, *51*, 355−431.
(2) Lee, J. C.; Timasheff, S. N. *J. Biol. Chem.* **1981**, *256*, 7193−7201.
(3) Kirkwood, J. G.; Goldberg, R. J. *J. Chem. Phys.* **1950**, *18*, 54−57.
(4) Schellman, J. A. *Biopolymers* **1978**, *17*, 1305−1322.
(5) Lee, J. C.; Timasheff, S. N. *Biochemistry* **1974**, *13*, 3, 257−265.
(6) Gekko, K.; Timasheff, S. N. *Biochemistry* **1981**, *20*, 4667−4676.
(7) Gekko, K.; Timasheff, S. N. *Biochemistry* **1981**, *20*, 4677−4686.
(8) Poklar, N.; Petrovcic, N.; Oblak, M.; Vesnaver, G. *Protein Sci.* **1999**, *8*, 832−840.
(9) Courtenay, E. S.; Capp, M. W.; Anderson, C. F.; Record, M. T., Jr. *Biochemistry* **2000**, *39*, 4455−4471.
(10) Casassa, E. F.; Eisenberg, H. *Adv. Protein Chem.* **1964**, *19*, 287−395.
(11) Greene, R. F., Jr.; Pace, C. N. *J. Biol. Chem.* **1974**, *249*, 5388−5393.
(12) Record, M. T. Jr.; Zhang, W.; Anderson, C. F. *Adv. Protein Chem.* **1998**, *51*, 281−353.
(13) Wyman, J.; Gill, S. J. *Binding and Linkage*: *Functional Chemistry of Biological Macromolecules*; University Science Books: Mill Valley, CA, 1990.
(14) Schellman, J. A. *Biophys. Chem.* **2002**, *96*, 91−101.
(15) Liu, Y. F.; Bolen, D. W. *Biochemistry* **1995**, *34*, 4, 12884−12891.
(16) Tanford, C. *J. Am. Chem. Soc.* **1964**, *86*, 2050−2059.
(17) Bolen, D. W. Protein Stabilizaiton by Naturally Occurring Osmolytes. In *Protein Structure, Stability, and Folding*; Humana Press: Totowa, NJ, 2001.
(18) Anderson, C. F.; Record, M. T., Jr. *J. Phys. Chem.* **1993**, *97*, 7116−7126.
(19) Mills, P.; Anderson, C. F.; Record, M. T., Jr. *J. Phys. Chem.* **1986**, *90*, 6541−6548.
(20) Tang, K. E. S.; Bloomfield, V. A. *Biophys. J.* **2002**, *82*, 2876−2991.
(21) Zou, Q.; Bennion, B. J.; Daggett, V.; Murphy, K. P. *J. Am. Chem. Soc.* **2002**, *124*, 1192−1202.
(22) Bennion, B. J.; Daggett, V. *PNAS* **2003**, *100*, 5142−5147.
(23) Tirado-Rives, J.; Orozco, M.; Jorgensen, W. L. *Biochemistry* **1997**, *36*, 6, 7313−7329.
(24) Alonso, D. O. V.; Daggett, V. *J. Mol. Biol.* **1995**, *247*, 501−520.
(25) Caflisch, A.; Karplus, M. *Struct. Fold. Des.* **1999**, *7*, 477−488.
(26) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, W.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(27) Ha, S. N.; Giammona, A.; Field, M.; Brady, J. W. *Carbohydrate Res.* **1988**, *180*, 207−221.
(28) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33−38.
(29) Karplus, M.; McCammon, J. A. *Nat. Struct. Biol.* **2002**, *9*, 646−652.
(30) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564−569.
(31) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395−2417.
(32) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(33) Duffy, E. M.; Severance, D. L.; Jorgensen, W. L. *Isr. J. Chem.* **1993**, *33*, 323−330.
(34) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliand, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(35) Forsyth, W. R.; Antosiewicz, J. M.; Robertson, A. D. *Proteins* **2002**, *48*, 388−403.
(36) Edgecomb, S. P.; Murphy, K. P. *Proteins* **2002**, *49*, 1−6.
(37) Schellman, J. A. *Biophys. J.* **2003**, *85*, 108−125.
(38) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−744.
(39) Lin, T. Y.; Timasheff, S. N. *Biochemistry* 94, *33*, 12695−12701.
(40) Richmond, T. J. *J. Mol. Biol.* **1984**, *178*, 63−89.