Solvation Model Based on Order Parameters and a Fast Sampling Method for the Calculation of the Solvation Free Energies of Peptides

Chong Gu,[†] Steve Lustig,[‡] and Bernhardt L. Trout*,[†]

Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue E19-502B, Cambridge Massachusetts 02139, and The DuPont Company, Central Research & Development, Experimental Station, Wilmington, Delaware 19880-0356

Received: August 15, 2005; In Final Form: November 30, 2005

An analytical solvation model is proposed as a function of an order parameter, which represents the local arrangement of water molecules in the first solvation shell of peptide atoms. The model is combined with a fast sampling method, rotational isomeric state Monte Carlo, to sample efficiently the torsional degrees of freedom on a peptide backbone. This order parameter solvation model is shown to reproduce without ad hoc fitting parameters the solvation free energies of single amino acids and tripeptides with slightly better accuracy than the generalized Born model but with several orders of magnitude improvement in efficiency. This method is a potential candidate for efficiently and accurately tackling some important issues in biophysical chemistry that are related to solvation, for example, protein folding, ligand binding, etc. Our results also present fundamental new insights into solvation. Specifically, the local water geometry, represented in this work by a properly defined order parameter, carries the majority, if not all, of the energetic information of solvation, including solute−solvent interactions and solvent reorganization in the presence of the solute.

1. Introduction

Thirteen years ago, Rand gave a perspective on the importance of hydration on molecular assembly and protein catalysis.[1] Since the majority of biological processes take place in solution, the effects of water must be of fundamental importance in biology. Thus, deep understanding of solvation effects in biosystems is essential. Current experimental measurement techniques, such as osmotic stress[2] and far-infrared laser vibration−rotation tunneling spectroscopy,[3] have been valuable in addressing this problem but are limited in the information that they can provide. However, computer modeling, which describes the solvation effects directly from a molecular perspective, yields direct molecular understanding.

For the description of solvation effects in biosystems, including proteins and DNA, solvation models must be both accurate and efficient. Even though the accuracy and efficiency of various models depend on the systems that they describe, they can be generally classified as shown in Figure 1. On the opposite corners are simulations in vacuum and with explicit solvent molecules. Without considering solvation, simulations in vacuum are obviously the most efficient and act as the basis for building the implicit solvent models described below. Explicit solvent models are regarded as the most accurate ones. However, the presence of solvent molecules incurs a huge computational expense because the majority of computer time is spent on the calculation of solvent−solvent interactions. To fill in the gap between the two extremes, various solvation models with higher efficiencies than those of explicit solvent models and higher accuracies than those of vacuum models have been developed.

One of the strategies is to treat the solvent as a continuum medium and add the solvation effects as a correction to the vacuum simulation without the explicit appearance of solvent molecules. Several models are built on this idea, and they are usually called implicit solvent models. According to the way in which they are developed, implicit solvent models generally fall into two categories.

In the first category, a linear relationship is assumed between the geometry of different groups in the solute molecule and their contribution to the solvation free energy. Included are the atom- or group-based solvent-accessible surface area (SASA) model,[4−8] the appropriately defined first solvation shell (FSS) model,[9] and the group contact model,[10] among which, the first two models are more computationally expensive because of the calculation of exposed surface area or volume. In fact, the SASA model may be the most expensive, because computing the second derivative with respect to the accessible surface area is time-consuming. The computational time may sometimes even be nearly equal to that of explicit solvent calculation.[11] Another severe defect of these models is the omission of solvent screening effects between charges. For this reason, the application of this type of model is usually confined only to hydrophobic hydration.

The second category of models is generally developed to treat the electrostatic screening. One most direct and simple way is to use Coulomb's law but to make the dielectric constant vary with the distance between charges (DDD). This approach was used frequently in the early days of molecular dynamics (MD).[12] Another representation is the Debye−Hückel (DH) model.[13] The computation time of these two models adds almost nothing to the total cost. However, they treat all pairs of charges identically regardless of the environment. Improvements were made in the effective energy function (EFF1),[14] which assumes the solvation free energy to be a Gaussian function of the dimensionless distance from the atom, the computation of which is only 50%

* Author to whom correspondence should be addressed. Phone: (617) 258-5021. Fax: (617) 258-5042. E-mail: trout@mit.edu.
[†] Massachusetts Institute of Technology.
[‡] Dupont Company.

10.1021/jp054602m CCC: $33.50
Published on Web 01/04/2006

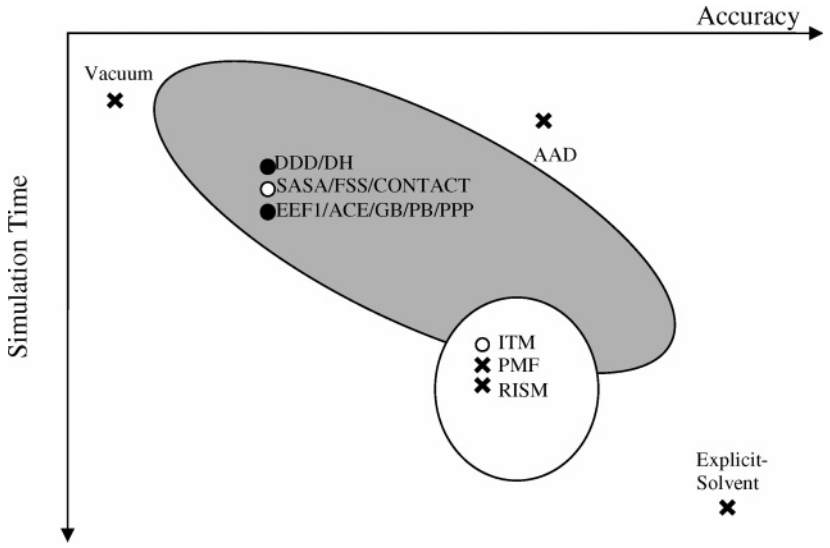

Figure 1. Schematic representation of the accuracy and efficiency of various solvation models. Open circles represent models that are designed for nonpolar solvation; filled circles represent models of electrostatic screening; crosses are models for both nonpolar and electrostatic screening interactions. The gray ellipse shows the range spanned by the implicit solvent models, and the white ellipse shows the range of the models on the basis of statistical mechanical formulas of second- and higher-order correlation functions.

slower than vacuum computations. The Poisson−Boltzmann (PB) equation and its analytical versions model the solvation effects in a more realistic way. The idea was poineered by Born in 1920 to calculate the hydration free energy of spherical ions,[15] extended by Kirkwood to treat arbitrary charge distributions in a spherical cavity,[16] and further improved such that the PB equation is solved analytically for simple boundary shapes[17] or numerically with finite-difference algorithms to treat the solute molecules with arbitrary shape.[18] However, the expensive computation of the finite-difference algorithm prevents the PB model from being efficiently incorporated into molecular mechanics calculations. To overcome the deficiency, various analytical continuum electrostatic potentials have been developed in recent years, including analytic continuum electrostatics (ACE)[11] and several versions of the generalized Born (GB) model.[19−23] The general idea behind these analytical models, based on the Born equation, is to propose an empirical parametrization of the effective Born radius, which represents neighbor charge effects on the central atom. Different from the total implicit treatment of the solvent, a related semiimplicit model called "polarizable pseudo-particles" (PPPs) was recently reported.[24] PPPs can be also used to calculate the electrostatic contribution of the solvation free energy. By discretization of the solvent region with regular grids and regarding each grid as a pseudo-particle represented by a dipole, the model is able to give a higher resolution of solvent structure but increases the computational expense.

A different strategy is to implement the reference interaction site model (RISM)[25] and its extended version for polar and ionic molecular systems, (XRISM),[26,27] which are developed on the basis of the statistical mechanics of molecular fluids. By solution of a site−site Ornstein−Zernike (SSOZ) equation with a closure of the hypernetted-chain (HNC) equation, RISM is able to give a detailed representation of the solvation structure, the pair distribution function, and to use it further to calculate the solvation free energy. Because of the inclusion of solvation structure, which is a great improvement on the implicit solvent models, RISM is able to give a more accurate estimation of the solvation free energy of a solute molecule with fixed structure in infinite dilution. Recent work also combines RISM with Monte Carlo simulated annealing to study the conformational stability of proteins in solution.[28−30] However, the solution of the SSOZ equation is still time-consuming, especially for three-dimensional systems, unless a very good initial density distribution function is given and a corresponding superior minimization scheme is implemented. To reduce the computational time, the direct correlation function used in RISM is generally simplified by using the potential of mean force (PMF),[31,32] where only two types of direct correlation functions are used, i.e., methane for nonpolar solute groups and water for polar solute groups. The two direct correlation functions are precalculated with an explicit solvent model, tabulated, and incorporated in the simulation to calculate the pair distribution function. By sacrifice of some accuracy in this way, a huge amount of computational time is saved. Another relevant model, recently developed, is the information theory model (ITM),[33] which was developed to investigate the temperature and pressure dependence of hydrophobic hydration. Since the pair distribution function is needed from either experiment or molecular simulation before the ITM computation, it is likely somewhat time-consuming. Moreover, ITM is accurate only for studying hydrophobic interactions.

Figure 1 presents a schematic of where different solvation models fit in the accuracy−simulation time coordinate. A good solvation model should be located as close as possible to the upper right corner, indicating a better accuracy and shorter simulation time. In this work, we present a simple analytical solvation model, called atom-based angle difference (AAD), which has these properties. AAD is based on a parametrization of the free energy of solvation as a function of a single order parameter, which represents the local water arrangement. To our knowledge, this model provides the first attempt at combining concepts of density functional theory with the calculation of peptide solvation, which is a completely new approach. Currently, the model focuses on small peptides, for which most of the molecule is solvated. It will be further generalized for larger molecules, such as proteins, in future work, following the same definition of the order parameter and correlation to the solvation free energy. This model is implemented with the assistance of a fast sampling method, rotational isomeric state Monte Carlo (RISMC).[13] In section 2, we describe the definition of the order parameter, AAD, used in this work. Another important idea in this work, the RISMC sampling method, is introduced in section 3, together with the other details of the

modeling. Parametrization of the solvation model is introduced at the beginning of section 4, followed by further validation through the computation of the solvation free energy of additional systems and the comparison with other solvation models. A discussion on the newly proposed method, as a whole, is provided at the end of section 4. Concluding remarks are given in section 5.

2. Atom-Based Angle Difference: An Order Parameter Representing the Local Water Arrangement and Overall Approach

Unlike the above-mentioned implicit solvent models, which are built on the basis of the solute and solvent interactions, we change the focus of our model to the local water arrangement. The essence of this idea comes from density functional theory: Given an external force field, the grand canonical free energy of the fluid system is a unique functional of the one particle density distribution, and the equilibrated density distribution can be obtained from the minimization of the grand canonical free energy functional. In the peptide solvation system, a given solute conformation can be regarded as an external force field to the solvent, no matter whether the interaction is van der Waals or electrostatic, etc. This interaction will result in a specific local water geometry and a corresponding solvation free energy. The local water geometry is a unique representation of the solute−solvent interaction, and it also includes the water rearrangement corresponding to the presence of the solute ignored in the other implicit solvent models. A similar idea was presented in an opposite way by Barciszewsk and co-workers,[34] who concluded in their work that structures of the solute are mainly affected by the three-dimensional arrangements of water. Also, as demonstrated in the most recent work of Raschke and Levitt,[35] the hydrophobic effects are closely related to the structural changes of the surrounding water caused by the appearance of a nonpolar solute. In this work, because of the complexity of the peptide solution, we do not intend to derive rigorously the free energy functional, but instead, we empirically relate the solvation free energy to a function of the local water geometry.

Thus, an order parameter is needed to represent the local water arrangement. This order parameter should fulfill at least two requirements: It should include both the number density and the orientation of the solvent molecules, and there should be a one-to-one mapping between the order parameter and the solvation free energy. Therefore, several obvious candidates related to hydrogen bonds are not acceptable, such as the average number of hydrogen bonds between the solute and the solvent, the hydration number, and the residence time of water molecules. The three-dimensional bond orientational order parameters introduced by Steinhardt et al.[36] may be acceptable choices. However, while they can differentiate systems with large differences in order, such as water in liquid and clathrate phases,[37] we found them unable to describe differences in water structure around solutes. For peptide systems, fluctuations in the order parameter overwhelm the difference between different values. Another possible choice of the order parameter is the tetrahedral order parameter, introduced by Chau and Hardwick to describe the network structure of water.[38] It was also later successfully applied to bulk water[39−41] and water in the CO_2 hydrate clathrates.[37] However, different from these topologically simple systems, the diversity of peptide atoms in solutions introduces a great geometric complexity with respect to the neighboring water structures and makes it difficult to implement directly the tetrahedral order parameter into the simulation. Therefore, we introduce a new order parameter, called the atom-based angle difference (AAD).

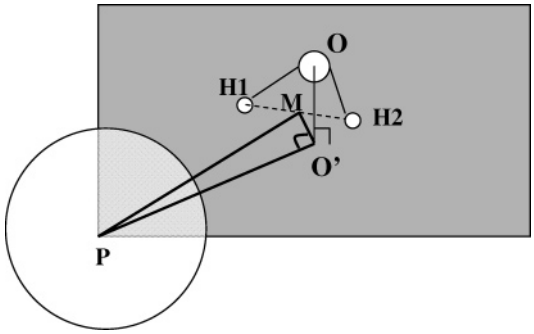

Figure 2. Definition of water orientation.

AAD is similar to the tetrahedral order parameter but differs from it by using the peptide atoms, not the water oxygen, as the central atom. AAD is expressed in eq 1

$$\mathrm{AAD}_i = (\cos \bar{\theta}_i - \cos \bar{\theta}_0)^2 \times \bar{N}_i \qquad (1)$$

where, $\bar{N}_i$, as referenced by the position of the oxygen atom, is the average number of waters of hydration within a cutoff distance from the center of the peptide atom i, namely, the first solvation shell, the range of which is defined as the van der Waals radius of the peptide atom plus 1.7 Å. This number represents half of the distance from the center of mass to the end of the first peak in the oxygen−oxygen pair distribution function in bulk water at 25 °C. θ_i is the angle that represents the orientation of water molecules relative to the central atom. θ_0 is the angle of a water molecule around a neutral central atom. Explicit descriptions of the two angles will be given in the following paragraphs.

As shown in Figure 2, O′ is the normal projection of O, the position of the water oxygen, onto the plane spanned by vectors PH1 and PH2, where P is the position of the peptide atom, and H1 and H2 are those of the two hydrogen atoms on water. Angle θ is the angle spanned by O′M and O′P, where M is the midpoint between H1 and H2. This definition of θ, although appearing complex, is, we think, the simplest representation that will be robust for different situations. From this definition, if O′ is on the line of MP and stands between the two points, then θ equals π; if O′ is on the line of MP but located on the opposite site of M relative to P, then θ is zero; for any other cases as shown in Figure 2, the value of θ is greater than zero and smaller than pi. Therefore, the larger the value θ, the closer the oxygen atom to the solute surface, while a smaller value of θ signifies that the hydrogen atom is closer.

Figure 3a shows three different distributions of the water orientation defined in Figure 2, which are collected and averaged from MD trajectories obtained using the CHARMM program package.[42] Figures 3b−d show the corresponding snapshots, in which the central spheres represent a hydrogen atom on the methyl group on an alanine side chain, a hydrogen atom on the amide group on a lysine side chain, and an oxygen atom on the carboxyl group on an aspartate side chain, respectively. In Figure 3a, squares stand for the angles around one of the hydrogen atoms on the methyl group on an alanine side chain. Because the alanine side chain is uncharged and hydrophobic, peaks in the distribution lie at both the small and the large angles, which means both hydrogen and oxygen on water molecules can to be close to the nonpolar solute surface, forming hydrogen bonds to one another. However, as shown by the curve, the angles near 180° are preferred around the amide hydrogen atom on a lysine side chain, which results from the strong attraction between the protonated amide group and the oxygen atoms on

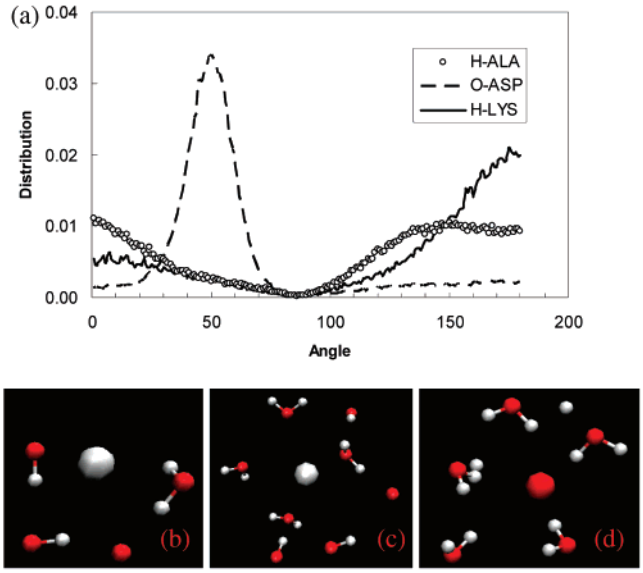

Figure 3. (a) Distributions of water orientations around the peptide atoms. Spheres are angles around one hydrogen atom on an alanine side chain; straight lines are those around the amide hydrogen on a lysine side chain; dashed lines refer to angles around the carboxyl oxygen on an aspartate side chain. (b) Snapshot of water around the hydrogen atom on the alanine side chain. (c) Snapshot of water around the amide hydrogen on the lysine side chain. (d) Snapshot of water around the carboxyl oxygen on the aspartate side chain.

water molecules. However, the distribution shown by crosses reveals a preferential binding between the negatively charged carboxyl oxygen atom on an aspartate side chain and the hydrogen atoms in water. Actually, besides these three examples, the angle distributions around different central atoms are all dissimilar to each other, which can be regarded as a characteristic feature of atoms on molecules in solution.

The average angle $\bar{\theta}_i$ in eq 1 can be determined from the distributions shown in Figure 3 in the following way

$$\bar{\theta}_i = \frac{\int_0^{180} \rho\theta \, d\theta}{\int_0^{180} \rho \, d\theta} \tag{2}$$

where ρ is the angle distribution. $\bar{\theta}_i$ is a measure of the interaction of the dipole moment of the water molecule with the atom in the peptide. The stronger the interaction between the atom and the surrounding water molecules, the more favorable the solvation. To be able to better discriminate different interactions, we subtract the cos $\bar{\theta}_0$ from the cos $\bar{\theta}_i$, where $\bar{\theta}_0$ is the average angle around the relative neutral alkane hydrogen atoms. The angle difference of each atom is then weighted by the number of water molecules around it, yielding the AAD, expressed in eq 1. AAD is a measure of the interaction of each atom with water and is thus added into the force field along with other parameters, for example, the van der Waals radii and charges.

3. Rotational Isomeric State Monte Carlo Sampling and Molecular Modeling Details for Computing the Solvation Free Energy

In the molecular modeling of solvation free energy in this work, three models, two sampling methods, and two formulas of free energy are used. One of the models is the order parameter solvation model, AAD, developed in this work, which will be explained in detail in the next section. The other two models are the explicit solvent and the GB models, which are two of the most frequently used models. To calculate the solvation free energy via the newly proposed model, the definition of the Helmholtz free energy as a function of the partition function is used together with RISMC sampling. The explicit solvent model and the GB model[23] are calculated by the CHARMM program package, in which a different scheme, thermodynamic integration with molecular dynamics sampling, is applied.

Rotational isomeric state (RIS) theory, as initiated by Flory,[43] has been extensively used to predict the mechanical and statistical mechanical properties of polymers by sampling the underlying polymer chain conformation. The fundamental idea of this approach is that a single polymer chain is limited to a discrete number of possible torsional angles corresponding to the minimum energy states. On the basis of this idea, any statistical methodology may then be applied to sample the chain conformations and therefore compute the properties. RISMC is a combination of RIS theory and Monte Carlo sampling and is a good candidate for sampling of peptide conformations. As an efficient sampling method, RISMC has been applied for designing polypeptides and polyelectrolytes[13] that can selectively recognize nanostructured substrates and has been validated by comparing results using it with experimental phage display observations. In RISMC as implemented in this work, peptides are described as continuous rotational isomeric state chains[13] in which bond lengths and bond angles are fixed at equilibrium values and only torsional angles are able to be changed. The peptide $\Phi-\Psi$ potential energy surface is generated by averaging the potential energies of the 20 amino acids with neutral caps over different side chain conformations at each fixed rotational angle pair around neighboring $N-C_\alpha$ and $C_\alpha-C'$ bonds. The potential energy includes the torsional energies of the backbone and the side chain as well as the van der Waals and electrostatic interactions between each atom within the same residue. The $\Phi-\Psi$ potential energies as a function of the pairwise-conditional rotational angles are tabulated and incorporated into the RISMC code. In the application of RISMC, a pair of $\Phi-\Psi$ angles is sampled within a continuous range from the rectangular tiles around the minima in the potential energy surface by a Monte Carlo method. The intramolecular energies are then obtained directly by interpolation from the energies around the selected point on the potential energy surface. The nonbond intermolecular interactions between atoms on different residues and the solvation terms are calculated by the corresponding models.

The Helmholtz free energy is then computed via

$$\Delta F_{solv} = F_{water} - F_{gas} = -kT \ln(Q_{water}/Q_{gas}) \tag{3}$$

where F_{water} and F_{gas} are the Helmholtz free energies of the solution and gas phases. F_{water} differs from F_{gas} by a solvation contribution as a function of the order parameter AAD. Q_{water} and Q_{gas} are the partition functions of the solution and gas phases, respectively.

Differently, in the CHARMM program package, the solvation free energy is calculated by thermodynamic integration, combined with molecular dynamics sampling of the peptide solution, either with the explicit solvent model or with the GB model. Thermodynamic integration introduces a number of intermediate states between the reactant peptide A and the product peptide B by constructing a linear combination of the two end states with a coupling parameter λ. The potential is expressed in eq 4

$$U(q_A,q_B,q_{env},\lambda) = (1-\lambda)U_A(q_A,q_{env}) + \lambda U_B(q_B,q_{env}) + U_{env}(q_{env}) \quad (4)$$

The free energy calculated by thermodynamic integration is

$$\Delta F_i = \int_0^1 \langle dU(\lambda)/d\lambda \rangle_{\lambda'} \, d\lambda' = \int_0^1 \langle U_B - U_A \rangle_{\lambda'} \, d\lambda' \quad (5)$$

where $\langle ... \rangle_{\lambda'}$ stands for the ensemble average over the system corresponding to the potential $U(\lambda)$at the given value of λ' and i is either the solution or the gas phase. A series of "windows" with fixed λ values are simulated, and the resulting free energy derivatives are integrated numerically to obtain the solvation free energy as shown in eq 6

$$\Delta F_{solv} = \Delta F_{water} - \Delta F_{gas} \quad (6)$$

The GB model, an analytical approximation of the Poisson−Boltzmann (PB) equation, is much more computationally efficient and can be directly implemented into molecular mechanics calculations. One of the most popular GB models was developed by Qiu and co-workers[44] and was later parametrized for implementation with the CHARMM force field.[45] The GB model has many successors, in which the Born radii in the PB model are reproduced. Evaluation of the Born radii is usually associated with two main approximations, the definition of the molecular volume or surface and the Coulomb field approximation (CFA). The first approximation involves both analytical[11,44−47] and nonanalytical[19,22,48] approaches to calculate the Born radii. The analytical methods are easily differentiable and are thus applicable to the energy minimization and dynamics. However, they usually underpredict the Born radii.[49,50] However, the nonanalytical approaches describe the molecular surface or volume by numerical grids, which are able to give more accurate predictions but are usually much more computationally costly. The CFA is designed to give the best performance for a charge at the center of spherical solute. However, it is not suitable for nonspherical solutes and charges near the solute surface. Therefore, the corrections to CFA are proposed for both the analytical and the nonanalytical approaches.[21] Unfortunately, among all these GB models, only one is programmed into the CHARMM package and can be combined with free energy calculations. This is the analytical approach without CFA corrections and is the one used in this work.[45] Also, in this version of the GB model, the nonpolar term, as represented by solvent-accessible surface area, is not considered. The results might be more accurate if the nonpolar term were included. We note, however, that this method should perform better than the other implicit solvent models.[51] Moreover, more accurate GB models will be much more computationally expensive.

The purpose of the MD calculation with explicit solvent is twofold. (1) Trajectories obtained from MD calculation can be applied to construct the solvation model, and (2) MD calculations can be used to test the validity of the newly proposed model. Therefore, good accuracy of the MD calculation with explicit solvent is crucial. To test this, the solvation free energy of the side chain of tyrosine was calculated in a system consisting of 1 solute molecule and 700 water molecules. The simulation is carried out in a constant temperature and pressure ensemble with a 200 ps equilibration run and 1 ns production run. Ten λ windows were used in the thermodynamic integration calculation. A solvation free energy of -3.90 ± 0.26 kcal/mol was obtained. This result matches well with -3.87 ± 0.04 kcal/mol, obtained in Shirts and co-workers[52] work, the latter number being more precise because it was obtained using a much more expensive computational approach: a system with 900 water molecules, 5 ns production run, and 62 λ windows.

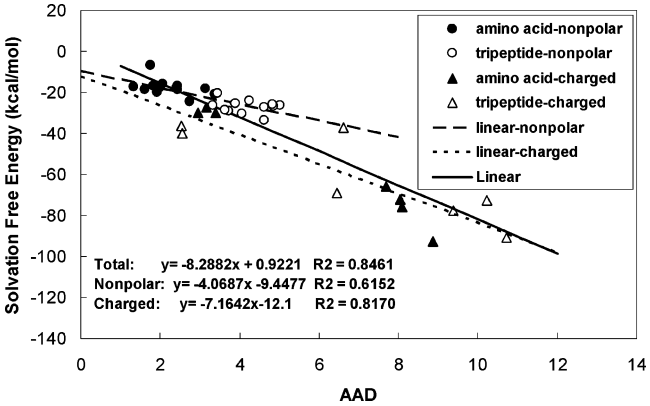

Figure 4. Relationship between AAD and solvation free energy.

Solvation free energies of 20 single amino acids (CH_3CO-X-NH_2) and 20 tripeptides with neutral terminals (CH_3CO-AXA-NH_2) are calculated for purpose of parametrization of the solvation model. To further validate the model, solvation free energies of 20 tripeptides with charged terminals (^+H_3N-AXA-COO^-) and 11 tripeptides with neutral terminals (CH_3CO-XXX-NH_2) are predicted by RISMC with the solvation model and compared with results obtained from MD sampling with the explicit solvent model and the GB model. Chloride and sodium are added as counterions into the system containing positively charged (lysine and arginine) and negatively charged (asparate and glutamate) amino acids, respectively. The positions of the counterions are randomly sampled around the peptide within the range of the maximum dimension of the peptide. The CHARMM22 force field is used in all of the calculations in this work. Free energies are obtained from a sample of 10^7 conformations for RISMC with the AAD model and 10^6 for MD with the GB model. Numbers computed via MD with the explicit solvent model are obtained from a sample of 0.5×10^6 conformations. Approximately 1000−2000 water molecules are used in the model, and the Particle Mesh Ewald algorithm is applied to treat electrostatic interactions. RISMC sampling with AAD and MD sampling in the gas phase use a cutoff of 10 Å, and the GB model is applied without a cutoff. MD samplings have a time step of 0.002 ps in a constant pressure and constant temperature ensemble. All of the calculations are at 298 K. Ten windows are used in thermodynamic integration computations of each solvation free energy, which is a good balance between accuracy and efficiency.

4. Solvation Energy as a Function of the Order Parameter

4.1. Linear Relationship between AAD and Solvation Free Energy. The total energy of the solvation phase E_t is given by the sum of conformational energy E_c and the solvation energy E_s

$$E_t = E_c + E_s(\mathrm{AAD}) \quad (7)$$

The first term on the right-hand side consists of the electrostatic energy, the van der Waals energy, and the torsional energy of the side chains and backbones. The second term represents the contributions of solute−solvent interaction and solvent reorganization energy, which is expressed as a function of the order parameter as described in more detail below.

In Figure 4, the solvation free energy is plotted as a function of AAD for each single amino acid and tripeptide with neutral

termini. All of the data are calculated by MD sampling with the explicit solvent model. The AADs are computed by summing the AAD of each atom in a given residue to give the AAD for that residue. The standard deviations of the results for a fit for all of the 40 peptides, as shown by the first equation in the figure, are 6.64 kcal/mol for the 20 single amino acids and 7.41 kcal/mol for the 20 tripeptides, respectively. Thus, the sum of AAD has a reasonably good linear relationship with the solvation free energy. To evaluate the AAD, the 40 data are divided into two groups and are correlated separately by lines. The first group includes the nonpolar and some of the polar residues, G, A, V, L, I, P, S, T, C, M, F, Y, and W, and the second group includes the more polar and charged residues, N, Q, K, R, H, D, and E. It turns out that the linear fit for the first group is not as good as that for the second group, because both the absolute value and the difference between the AAD for the nonpolar peptides are smaller, and therefore the values are easily affected by computational errors. Also, it is possible that the order parameter may not treat the atoms in the nonpolar group as well as in the polar groups. However, as shown in the solvation model in the following section, the deviation from the linear fit can be adjusted by the geometry correction term in the solvation model.

4.2. Parametrization of the Solvation Model. In Figure 4, the AADs are calculated directly from the MD sampling trajectories of each individual peptide. However, for the prediction of the solvation free energies for other peptides, we need to invert the problem, predicting the order parameter for a given residue in any environment, including the influence of neighboring atoms. This will lead to a solvation model that can be incorporated directly into the Hamiltonian. We choose the form in eq 8

$$E_s(\mathrm{AAD}) = K_1 \sum_{\substack{i=\text{peptide}\\ \text{atoms}}} \mathrm{AAD}_i \left(1 - K_2 \sum_{\substack{j=\text{nonbonded}\\ \text{neighbors}}} \frac{V_j}{r_{ij}^4}\right) + K_3 \quad (8)$$

where K_1, K_2, and K_3 are solvation parameters, which are the same for all the peptides and need to be optimized. The summation of AAD_i is over all the peptide atoms, including the backbone, the side chains, and the counterions. The summation of the term V_j/r_{ij}^4 is over the nonbonded neighbors of atom i. V_j is the volume of each neighboring atom calculated by the van der Waals radius. r_{ij} is the distance between the central atom and its neighbor. The V_j/r_{ij}^4 term follows the definition of the polarization energy in the original GB equation.[44] It represents the loss of free energy caused by a classical charge-induced dipole interaction between the charge on atom i and the dielectric medium displaced by atom j. However, different from the original equation, which has three V_j/r_{ij}^4 terms representing the bond, angle, and nonbond effects, only the influence of nonbonded neighbors is considered here. This is because the major effects contributed by the bond and angle interactions and part of the nonbonded interactions are already considered in the values of the AAD, and only a minor correction is needed to further include the geometry correction. The so-called nonbonded neighbor in the second summation term in eq 8 is counted without a cutoff distance for the current application on small peptides. However, for a future extension of the model to globular proteins, a cutoff distance will be needed, and the model will be developed carefully with the shifted or switched potential schemes to avoid the discontinuity problem. The AAD_i of each atom in eq 8 is calculated and averaged over MD trajectories of 20 amino acids and 20 tripeptides with neutral termini. All of the atoms are divided into 36 categories according to their AAD values, which distinguish their chemical properties. The terminologies in the lettered rows are similar to the ones used in the force field,[42] in which the first letter stands for the type of atom and the latter ones represent the environment the atoms are in; for example, "CA" means the α-carbon on the peptide backbone. In comparison to the CHARMM22 force field, in our solvation model, there are only three different definitions, which are "CD" for the carboxyl carbon on aspartate and glutamate, "HA2" for hydrogen on the CH_2 group on argnine and lysine, and "HA3" for hydrogen on the CH_2 group on aspartate and glutamate. The deviation from the average AADs for each category is no more than 10%. The values are shown in Table 1.

TABLE 1: Values of AAD

C	CA	CC	CD	CP2	CP3	CPH1	CPH2	CPT
0.0000	0.0449	0.0000	0.0067	0.0000	0.0000	0.0129	0.0384	0.1362
CT1	CT2	CT3	CY	H	HA	HA2	HA3	NH2
0.0000	0.0000	0.0001	0.0066	0.0332	0.0000	0.0217	0.0159	0.0032
NH3	NR1	NR2	NY	O	OC	OH1	S	HB
0.0000	0.0139	0.6147	0.1769	0.9021	1.6294	0.4280	0.1235	0.0012
HC	HP	HR1	HR3	HS	NC2	NH1	NA	CLA
0.1376	0.0010	0.0110	0.0049	0.0450	0.0084	0.0000	1.9348	5.6305

The three parameters in eq 8 are fitted, and the final result is expressed in eq 9

$$E_s(\mathrm{AAD}) = -14.2556 \sum_{\substack{i=\text{peptide}\\ \text{atoms}}} \mathrm{AAD}_i \left(1 - 0.1731 \sum_{\substack{j=\text{nonbonded}\\ \text{neighbors}}} \frac{V_j}{r_{ij}^4}\right) - 3.1755 \quad (9)$$

Note that, as shown in Table 1, the values of AAD_i are averaged from the trajectories calculated by the explicit solvent model. Therefore, they include both the self-energy of the atom and screening effects on the electrostatic interactions between the partial charges. Moreover, the negative coefficient of the V_j/r_{ij}^4 term also includes the screening effects between neighboring partial charges. As shown in Table 1, the values of the AAD for hydrophilic atoms are generally larger than those for hydrophobic ones. Therefore, when another atom comes close to a hydrophilic atom, the penalty for the solvation free energy is much greater than if it comes close to a hydrophobic atom. That means that hydrophilic atoms need to be separated far away from each other to have a lower solvation free energy, while the hydrophobic ones do not. Therefore, the current single coefficient in the geometry correction term for nonbonded interactions is adequate to treat the hydrophobic and hydrophilic atoms. However, to further improve the accuracy of the model, geometry corrections between different types of atoms, including the nonpolar atoms, the polar atoms, and the charged atoms with the same and different signs could be treated differently in future work.

The parameters are fitted by the Levenberg−Marquardt method with 10^6 samplings on each of the 20 single amino acids (CH_3CO-X-NH_2) and 20 tripeptides (CH_3CO-AXA-NH_2). The χ^2 merit function is expressed as

$$\chi^2 = \sum_{i=1}^{40} [F_{\mathrm{water},i} - F_{\mathrm{water}}(\mathrm{AAD},\mathbf{K})]^2 \quad (10)$$

TABLE 2: Summary of the Efficiency and Accuracy

	standard deviation (kcal/mol)				
	CH_3CO-X-NH_2	CH_3CO-AXA-NH_2	^+H_3N-AXA-COO^-	CH_3CO-XXX-NH_2	CPU time[a]
AAD + RISMC	6.14	6.09	7.81	5.05	25 min
GB+MD	5.84	8.31	12.16	6.25	8 h
explicit solvent + MD[b]					700 h

[a] CPU time refers to 10^6 samplings of CH_3CO-MIT-NH_2. [b] Results calculated by the explicit solvent model with MD sampling are regarded as the standards, and standard deviations of all the other results are calculated relative to those obtained by this method.

where $F_{water,i}$ represents the free energy of each peptide in water by MD sampling with an explicit solvent model and F_{water}(AAD,**K**) is the corresponding result obtained from RISMC sampling with the AAD model. **K** is the vector consisting of the three parameters in eq 8 that are needed to be fit. As a result, χ^2 equals 4742.60, which corresponds to an average standard deviation of 10.88 kcal/mol for each peptide. Notice that this number is a little bit larger than those shown in Table 2, because the standard deviation obtained here is of the absolute free energy in water while the ones in Table 2 are related to the solvation free energies. The derivatives of χ^2 with respect to each parameter are

$$\beta_1 = \frac{\partial \chi^2}{\partial K_1} = 1591.29$$

$$\beta_2 = \frac{\partial \chi^2}{\partial K_2} = 69865.10$$

$$\beta_3 = \frac{\partial \chi^2}{\partial K_3} = 1598.40 \qquad (11)$$

As indicated in eq 11, K_2, which is closely related to the conformational movement of the solute atoms, is the most sensitive parameter among the three. To estimate the sensitivity of each parameter in a more direct way, changing each parameter by 1 will cause an average deviation of free energy in water for each residue of about 6.31 kcal/mol for K_1, 41.79 kcal/mol for K_2, and 6.32 kcal/mol for K_3.

The optimized results of the solvation free energies of 20 single amino acids (CH_3CO-X-NH_2) and 20 tripeptides both with neutral termini (CH_3CO-AXA-NH_2), calculated via RISMC sampling together with the AAD solvation model are shown in Figure 5. It should be noted that by using eq 9 with the optimized parameters the calculated solvation free energies have smaller standard deviations compared to those shown in Figure 4, even with the generalized AADs. Results of this model are also compared with those from MD sampling of the GB solvation model believed to be the most accurate continuum solvation model other than the PB equation. The *x*-axis represents the results calculated by MD sampling of the explicit solvent system, and the *y*-axis shows those obtained from the other models. The straight line stands for the perfect fitting. The accuracies of these methods are given in Table 2.

4.3. Validation of the Solvation Model. Aside from the above optimized results, the newly proposed solvation energy model is further used to predict the solvation free energies for additional systems. Figure 6 shows predictions of the solvation free energies of the 20 tripeptides with charged terminals (^+H_3N-AXA-COO^-). The standard deviations for both methods are shown in Table 2.

The AAD model is also used to predict the relative solvation free energies, instead of the absolute solvation free energies, of a group of representative tripeptides and compared with the GB model. All of the relative solvation free energies are relative to

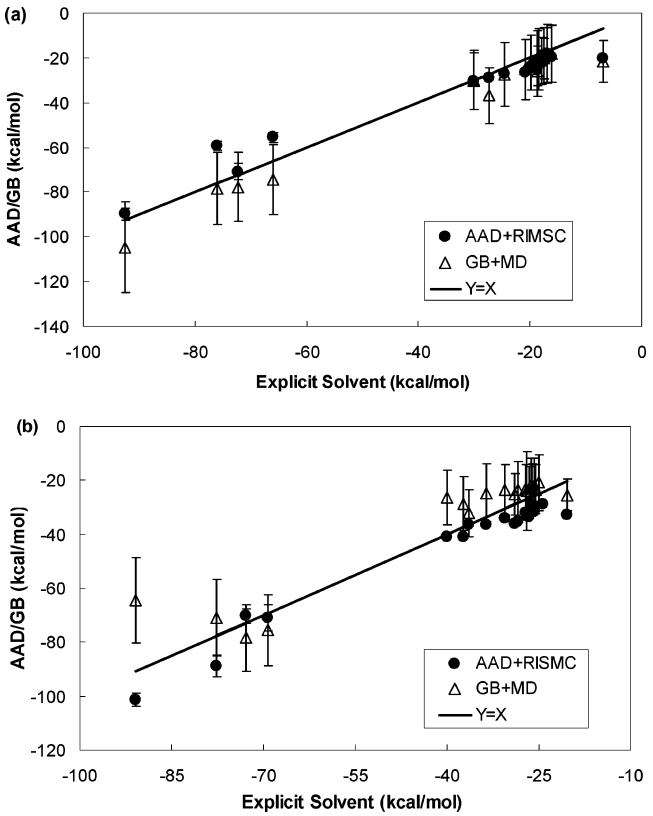

Figure 5. Solvation free energies optimized by RISMC sampling with the AAD solvation model and predicted by MD sampling of the GB model vs MD sampling with the explicit solvent model: (a) single amino acids with neutral termini; (b) tripeptides with neutral termini.

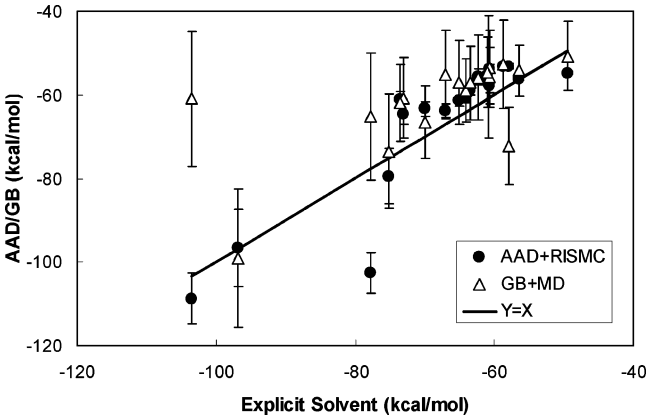

Figure 6. Comparison of the solvation free energies of tripeptides with charged termini predicted by the AAD and GB models vs results predicted by the explicit solvent model.

the corresponding solvation free energies of the tripeptides with the middle residues mutated to alanine. For example, the relative solvation free energy of "FIR" means the solvation free energy difference between "FIR" and "FAR". The errors of the results calculated by these two methods are shown in Figure 7.

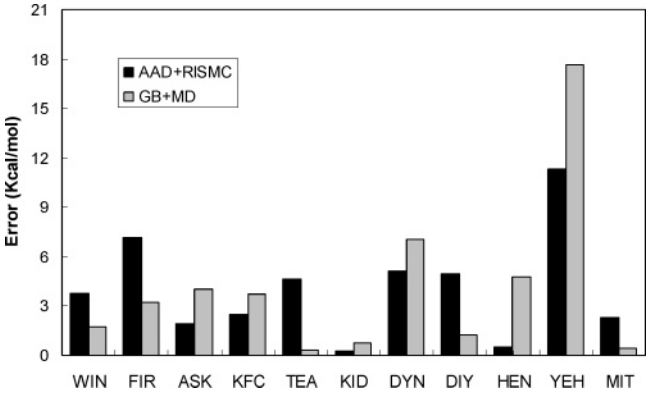

Figure 7. Comparison of the errors of the relative solvation free energies of tripeptides predicted by the AAD and GB models.

4.4. Discussion about the Method. Table 2 summarizes the standard deviations of the optimized and predicted results discussed above. The differences between our results and those from MD sampling on the explicit solvent model are caused by both the model and the sampling. From the model side, there are larger differences in the charged residues than the neutral ones, which means that the model is still not able to fully represent the strong charge interactions. This could be corrected by adding more terms, noting that our model is still better than the GB model. Also, the fitting results of the three parameters may not be exactly the most optimal ones because of the fluctuations of the free energies caused by MC samplings. A better fitting may be obtained by more samplings, which is, however, too expensive to compute. Moreover, one of the assumptions in the RIS theory is the fixed bond length and bond angle, which may cause some differences when sampling the configurational space, especially for peptides with strong electrostatic interactions. However, the problem may be less severe for longer peptides, because they are more flexible to be bent even with fixed bond lengths and angles. From the sampling side, for residues in different environments, RISMC is sampling on the same $\Phi-\Psi$ energy surface obtained from 20 single amino acids with neutral caps, disregarding the influence of the environment, through which it gains efficiency but slightly loses accuracy. Last but not least, 0.5×10^6 MD samplings with the explicit solvent model is a compromise between the computational capacity and the convergence, which may not be able to sample all of the configurational space. Despite these analyses, it is clearly shown in Table 2 that the AAD model is able to predict the solvation free energy as good as or even slightly better than the GB model. Moreover, with the assistance of RISMC sampling on the continuum $\Phi-\Psi$ energy surface and through the use of eq 3 to calculate the solvation free energy, the newly proposed method is much more rapid than MD sampling with either the GB model or the explicit solvent model.

The success of this order parameter solvation model is based on the following considerations. According to the density functional theory, the free energy of the system is a unique functional of the fluid density distribution, which indicated that the geometry of the solvent molecules carries the majority of the energetic information of solvation. From this base, our order parameter solvation model relates the solvation free energy empirically with an order parameter that represents the local water arrangement. However, the implicit solvent models only consider the geometry of the solute itself without taking into account the structural information of the solvent, which is the major factor that causes inaccuracy. Solvation models based on the statistical mechanical theories of inhomogeneous fluid consider the solvent structural information as our model does. However, this type of model generally regards the solute molecule as a whole and considers the range of solvent distributions around the solute molecule, which is not an efficient way for the computation of peptide solutions. For this reason, the original RISM is used only to calculate the solvent distribution around a fixed conformation of a simple solute molecule in infinite dilute solution. Even though the Monte Carlo algorithm is implemented in the later version of RISM to sample the conformational change of the solute molecule, the computation is even more time-consuming than the original one. However, in our model, only the solvent structures around the carefully defined first solvation shell of each atom are considered, which is proven by our result to carry most of the information of solvation. The geometric information is represented by an atom-based order parameter and is included as a new component into the force field. In this way, our order parameter solvation model introduces much more flexibility and efficiency in computation than the solvation models based on the inhomogeneous fluid theory.

In summary, by inclusion of an atom-based order parameter, which represents the local water arrangement, our model, with a concise parametrized form, is able to give a more accurate and efficient calculation of solvation.

5. Concluding Remarks

The simple analytical solvation model described above, using a sensitive order parameter, AAD, to represent the interaction between solute and solvent, including solvent reorganization, is able to give accurate and rapid prediction of peptide solvation free energies. For the systems studied in this work, the newly proposed model is demonstrated to give results as good as or even slightly better than those of the GB model but with a much simpler form. However, the validity of this model on larger peptides needs to be tested further. Furthermore, our model could be continually improved, particularly in its representation of strong electrostatic interactions. More significantly, by the inclusion of RISMC sampling, our new method is much more efficient than thermodynamic integration with the GB model. Because this AAD model is fully analytical, derivatives of the energy with respect to movements of the atoms are also available and allow the effects of solvation to be included efficiently in energy minimization and molecular dynamics, etc. Implementations of the order parameter solvation model together with the RISMC sampling on some real-world applications and comparison with experimental work are in progress.

Acknowledgment. The work is supported by the funding from the DuPont−MIT alliance.

References and Notes

(1) Rand, R. P. *Science* **1992**, *256*, 618.
(2) Parsegian, V. A.; Rand, R. P.; Fuller, N. L.; Rau, D. C. *Methods Enzymol.* **1986**, *127*, 400.
(3) Liu, K.; Cruzan, J.; Saykally, R. *Science* **1996**, *271*, 929.
(4) Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227.
(5) Eisenberg, D.; Mclachlan, A. D. *Nature* **1986**, *319*, 199.
(6) Ooi, T.; Oobatake, M. *Proc. Natl. Acad. Sci. U.S.A.* **1991**, *88*, 2859.
(7) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086.
(8) Schiffer, C. A.; Caldwell, J. W.; Stroud, R. M.; Kollman, P. A. *Protein Sci.* **1992**, *1*, 396.
(9) Kang, Y. K.; Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1987**, *91*, 4105.
(10) Colonnacesari, F.; Sander, C. *Biophys. J.* **1990**, *57*, 1103.
(11) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578.

(12) Weiner, S.; Kollman, P.; Case, D.; Singh, U.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. *J. Am. Chem. Soc.* **1984**, *106*, 765.
(13) Lustig, S. R.; Jagota, A. *Mater. Res. Soc.* **2002**, *724*, N4.6.1.
(14) Lazaridis, T.; Karplus, M. *Proteins* **1999**, *35*, 133.
(15) Born, M. Z. *Physics* **1920**, *1*, 45.
(16) Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351.
(17) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333.
(18) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K. *Proteins* **1986**, *1*, 47.
(19) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.
(20) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239.
(21) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *J. Chem. Phys.* **2002**, *116*, 10606.
(22) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983.
(23) Dominy, B. N.; Brooks, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765.
(24) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Comput. Chem.* **2004**, *25*, 1015.
(25) Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1972**, *57*, 1930.
(26) Yu, H. A.; Pettitt, B. M.; Karplus, M. *J. Am. Chem. Soc.* **1991**, *113*, 2425.
(27) Hirata, F.; Pettitt, B. M.; Rossky, P. J. *J. Chem. Phys.* **1982**, *77*, 509.
(28) Kinoshita, M.; Okamoto, Y.; Hirata, F. *J. Am. Chem. Soc.* **1998**, *120*, 1855.
(29) Kinoshita, M.; Okamoto, Y.; Hirata, F. *J. Chem. Phys.* **1999**, *110*, 4090.
(30) Mitsutake, A.; Kinoshita, M.; Okamoto, Y.; Hirata, F. *J. Phys. Chem. B* **2004**, *108*, 19002.
(31) Garde, S.; Hummer, G.; Garcia, A.; Pratt, L.; Paulaitis, M. *Phys. Rev. E* **1996**, *53*, R4310.
(32) Hummer, G.; Soumpasis, D. *Phys. Rev. E* **1994**, *50*, 5085.
(33) Hummer, G.; Garde, S.; Garcia, A.; Paulaitis, M.; Pratt, L. *J. Phys. Chem. B* **1998**, *102*, 10469.
(34) Barciszewski, J.; Jurczak, J.; Porowski, S.; Specht, T.; Erdmann, V. *Eur. J. Biochem.* **1999**, *260*, 293.
(35) Raschke, T. M.; Levitt, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6777.
(36) Steinhardt, P.; Nelson, D.; Ronchetti, M. *Phys. Rev. B* **1983**, *28*, 784.
(37) Radhakrishnan, R.; Trout, B. L. *J. Chem. Phys.* **2002**, *117*, 1786.
(38) Chau, P. L.; Hardwick, A. J. *Mol. Phys.* **1998**, *93*, 511.
(39) Errington, J. R.; Debenedetti, P. G. *Nature* **2001**, *409*, 318.
(40) Radhakrishnan, R.; Trout, B. L. *J. Am. Chem. Soc.* **2003**, *125*, 7743.
(41) Radhakrishnan, R.; Trout, B. L. *Phys. Rev. Lett.* **2003**, *90*.
(42) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586.
(43) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Hanser Publisher: New York, 1988.
(44) Qiu, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* **1997**, *101*, 3005.
(45) Dominy, B.; Brooks, C. *J. Phys. Chem. B* **1999**, *103*, 3765.
(46) Onufriev, A.; Case, D.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.
(47) Chambers, C.; Hawkins, G.; Cramer, C.; Truhlar, D. *J. Phys. Chem.* **1996**, *100*, 16385.
(48) Scarsi, M.; Apostolakis, J.; Caflisch, A. *J. Phys. Chem. A* **1997**, *101*, 8098.
(49) Swanson, J.; Mongan, J.; McCammon, J. *J. Phys. Chem. B* **2005**, *109*, 14769.
(50) Zhu, J.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2005**, *109*, 3008.
(51) Edinger, S.; Cortis, C.; Shenkin, P.; Friesner, R. *J. Phys. Chem. B* **1997**, *101*, 1190.
(52) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.