Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids

Brian N. Dominy and Charles L. Brooks, III*

Department of Molecular Biology, TPC6, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received: November 17, 1998

The generalized Born model proposed by Still and co-workers (Qui, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014) is parametrized specifically for proteins, peptides, and nucleic acids within the CHARMM all hydrogen and polar hydrogen force fields. A database of atomic electrostatic environments and molecular electrostatic solvation free energies comprising amino acid residues, nucleic acid bases, dipeptides, dinucleotides, proteins, and DNA strands is established, and numerical finite difference Poisson calculations are solved using atomic radii and charges from the corresponding force field parameters. These data provide the necessary input to parametrize generalized Born models for a particular force field. The performance of these models in reproducing overall molecular solvation trends is examined and found to be quite good. Furthermore, calculations of electrostatic solvation free energy differences suggest that conformational free energy changes are well reproduced. Finally, the utilization of this generalized Born model in molecular dynamics simulations (both with and without cutoffs) is shown to give excellent agreement with explicit solvent simulations of a small 56-residue protein in water.

I. Introduction

Molecular modeling and simulation are essential tools in the repertoire of theoretical and computational chemistry. Central to the success of these methods is the accurate representation of the material of focus and often the solvent environment. In fact, in studies of a biological solute in water, the water component overwhelmingly dominates the computation. This factor remains limiting and to date has hampered the quest for relevant studies of protein folding and dynamics on microsecond time scales by direct methods.[1]

One means to overcome this limitation on computation time is to replace the explicitly represented solvent by a modified set of interactions between the atoms that mimics the relevant features of the solvent of interest. Such ideas motivated early work by Kuntz and co-workers, who introduced an empirical, spatially dependent dielectric function to modify the electrostatic interactions between solute atoms of a biological molecule.[2] For water and other high dielectric polar fluids, electrostatic interactions dominate these solvent-mediated interactions. In these cases one may use continuum electrostatic theory as a cornerstone for the development of implicit solvent models. However, the explicit numerical solution of Poisson's equation is also too costly, from a computational perspective, to permit useful long time dynamics of biological molecules to be routinely studied.[3,4]

Nevertheless, this framework has been demonstrated to provide accurate representations of solvation and solvent-mediated interactions[5,6] and provides the reference for the development of approximate analytic solutions to the classical electrostatic equations. Such approximate continuum solvent models represent the solvent in terms of its average effect on a solute. Although the details of the implementation vary, the primary advantage of these methods is speed. In many cases, implicit solvent models are able to substitute for the effects of explicit solvent while increasing the speed of energy calculations substantially. These models can facilitate calculations that are intractable using atomic solvent models. This includes energy calculations on enormous databases of static structures as well as molecular dynamics simulations of larger systems and over long time scales.[7]

One model that provides an approximate continuum solvent representation is the generalized Born (GB) model introduced by Still and co-workers:[8]

$$G_{\mathrm{pol}} = -166\left(1 - \frac{1}{\epsilon}\right)\sum_i^N\sum_j^N \frac{q_i q_j}{\sqrt{(r_{ij}^2 + \alpha_i\alpha_j e^{-D_{ij}})}} \quad \text{with} \quad D_{ij} = \frac{r_{ij}^2}{4\alpha_i\alpha_j} \qquad (1)$$

Based on the Born model for ionic solvation,[9] the generalized Born model extends this formalism to treat solutes containing multiple charged particles and an arbitrarily shaped molecular surface. The generalized Born model requires a parameter for each atomic site within the system that is contingent on its extent of burial. A rapid, analytical approach to calculating these atomic parameters, or Born radii (α_i), was developed by Still and co-workers[10] and is utilized in this work.

We also note that a number of papers have recently been published describing a variety of continuum solvent models either based directly upon or closely related to the Born model. Some of these models focus primarily on the calculation of the electrostatic component of the solvation free energy[11−16] while others are parametrized to reproduce the total solvation free energy.[7] The majority of implicit solvation models have been parametrized and applied to small molecules and have been shown to accurately reproduce experimental solvation energies for these compounds.

The objective of this work is to extend the developments of Still and others and to establish a parametrization of Still's model that is consistent with the CHARMM force field and is capable

[*] Corresponding author. E-mail: brooks@scripps.edu, http://www-.scripps.edu/brooks.

10.1021/jp984440c CCC: $18.00
Published on Web 04/16/1999

of reproducing the electrostatic solvation free energies using parameters from this force field, i.e., atomic radii and charges, as calculated by the finite difference solution to the Poisson equation (FDPB).[17] It is possible to use methods similar to those employed here to optimize against experimental solvation free energies. However, these values are generally not available for macromolecules. For this reason, the FDPB calculation is used as the reference solvation energy in the work described below.

This paper extends the approach of Still and co-workers to macromolecular structures including proteins and nucleic acid strands. By using a set of molecules that includes proteins and nucleic acid strands as well as small molecules, we include a significant amount of diversity in atomic burial, leading to a robust fit to solvation energies for a range of molecular environments. We demonstrate that this parametrization is capable of accurately reproducing the solvation energies (relative to finite difference Poisson−Boltzmann calculations) for macromolecules of biological interest.

Our paper is organized as follows. The Methods section is divided into five subsections. We begin by describing the relevant equations we seek to optimize and the atomic and molecular databases we have established for this purpose. We then provide details of the extensive set of finite difference Poisson−Boltzmann calculations carried out to establish the target electrostatic solvation energies based on the database of atomic and molecular environments. Next, we present the fitting procedure used to optimize the Born radii and molecular solvation energies. Finally, we describe the protocols used to explore our implementation of GB in molecular dynamics simulations. In the Results and Discussion section of the paper, we present findings for the fitting of the atomic-based parameters to the established databases and the resulting performance of these parameters in reproducing the molecular electrostatic solvation energies of database structures. Next, we examine the correlation between changes in the electrostatic solvation free energy with conformation for a peptide and small protein. We then discuss the results of several nanosecond long molecular dynamics simulations on a small protein, segment B1 of streptococcal protein G,[18−20] and the behavior of our GB model against other simple "implicit" solvent models as well as an explicit solvated simulation. Finally we discuss the use of long range cutoffs in calculations of energies and forces for molecular simulations using our GB model. The paper concludes with an overview of our findings.

II. Methods

In this section we review the analytical approximation introduced by Still to compute the Born radius of an atom in a molecule.[10] We describe the composition and preparation of databases used in the parametrization of this empirical relationship, and we continue with a description of the calculation of the finite difference solution to the Poisson equation using DelPhi[21] to obtain atomic and molecular polarization free energies. The section concludes with a description of the protocol used in the molecular dynamics simulations of protein G.

II.A. Computing Born Radii and Electrostatic Solvation Energies. The empirical expression first proposed by Still relates the "generalized Born" radius of an atom in a specific molecular environment to the polarization energy of that atom in the same environment through the classical equation for the Born polarization free energy (eq 2)

$$\alpha_i = -166/G_{\text{pol},i} \tag{2}$$

The analytical formula for atomic polarization energy utilized by Still is based on the premise that the polarization of an external medium (such as water) due to a charge is reduced proportionally to the volume of displaced dielectric around the charge. This displaced dielectric originates from the volume occupied by atoms surrounding the point charge of interest. The reduction of polarization is inversely proportional to the distance (from the charge to the displaced volume elements) raised to the fourth power, as derived from the classical charge-induced dipole interaction.[22] It has also proven useful to separate the contributions to the burial of a specific point charge based on the chemical topology. In the model of Still, the relative contribution of surrounding atoms toward the reduction of polarization energy is partitioned based on topology as follows: atoms bonded to the charge of interest, atoms within a bond angle with the charge of interest, and atoms related to the charge only through nonbonded interactions.[10] We use a linearized form of Still's original empirical formula (eq 3).

$$G_{\text{pol},i} = \left(1 - \frac{1}{\epsilon}\right)\left[\frac{1}{\lambda}\left(\frac{-166}{R_{\text{vdW},i}}\right) + P_1\left(\frac{166}{R^2_{\text{vdW},i}}\right) + \sum_j^{\text{bond}} \frac{P_2 V_j}{r_{ij}^4} + \sum_j^{\text{angle}} \frac{P_3 V_j}{r_{ij}^4} + \sum_j^{\text{nonbond}} \frac{P_4 V_j}{r_{ij}^4}\text{CCF}\right]$$

where CCF =

$$\begin{cases} 1.0; & \left(\dfrac{r_{ij}}{R_{\text{vdW},i} + R_{\text{vdW},j}}\right)^2 > \dfrac{1}{P_5} \\ \left\{0.5\left[1.0 - \cos\left(\left(\dfrac{r_{ij}}{R_{\text{vdW},i} + R_{\text{vdW},j}}\right)^2 P_5 \pi\right)\right]\right\}^2; & \left(\dfrac{r_{ij}}{R_{\text{vdW},i} + R_{\text{vdW},j}}\right)^2 \le \dfrac{1}{P_5} \end{cases} \tag{3}$$

In this expression, r_{ij} is the distance from the point charge of interest, i, and an uncharged atom j in the molecule; $R_{\text{vdW},i}$ is the van der Waals radius of atom i (taken as the radius parameter $R_{\min}$ for atom i in the CHARMM parameters); ϵ is the dielectric constant of the solvent; and V_j is the atomic volume of atom j. We note that the atomic volume used in our formulation does not account for shared atomic volume as in the work by Qui et al. and is simply $4/3\pi R^3_{\text{vdW},i}$. The effective reduction in atomic volume due to neighboring atoms is absorbed to some extent by the fitting parameters. The fitting parameters are P_1, P_2, P_3, P_4, P_5, and λ. From this expression for the solvation energy of a unit charge ion, the Born radius is computed using the Born equation given above (eq 2).

A nonlinear version of eq 3 was fit in the original work of Still by minimizing the root-mean-square difference (RMSD) between the generalized Born polarization energy (solvation energy) and the finite difference Poisson−Boltzmann (FDPB) corrected reaction field over a collection of atoms. The search technique employed was a simulated annealing approach[23] operating in the solution space defined by the fitting parameters P_1−P_5. This algorithm does not guarantee the optimal solution in finite time and would be prohibitive for optimizing the large number of atoms within macromolecular systems. Our linearized form of the original equation permits multivariate linear regression to be performed over the space of P_1−P_4 and λ. The parameter P_5 is chosen by identifying the minimum in the target function versus P$_5$. This fitting method does guarantee the optimal solution for the linearized equation and is extremely rapid, which is important when considering the large number

TABLE 1: Structural Databases Utilized in This Study[a]

param19 composite			param22 composite		
database	no. molec.	no. atom	database	no. molec.	no. atom
single AA	20	203	single AA	20	324
di-AA	210	4263	di-AA	210	6804
proteins	22	11995	proteins	22	19417
			single NA	5	151
			di-NA	15	921
			strands	22	13496

[a] The param19 and param22 databases are comprised of three and six component databases, respectively. The number of molecules and atoms contained within each of these component databases is given.

of atoms in our macromolecular databases. In addition, the new linear function only differs from the nonlinear version by approximately 1% for the range of radii within the CHARMM parameter set. We have also introduced the parameter (λ) used to scale the van der Waals (vdW) radius of the atom in question. This parameter can be used to eliminate systematic error in the fitting of eq 3 above, as discussed later in this paper.

II.B. Structural Databases. As evident from the relationships defined in the previous section, the parametrization of this empirical form of the electrostatic solvation free energy (eqs 2 and 3) focuses most directly on the nature of atomic environments encountered by atoms in specific molecules. Therefore, the detailed charges used for these atoms, and expressed as a component of a particular force field, do not play a part in the parametrization of the generalized Born model. We believe this is a positive attribute of the particular empirical model proposed by Still that will permit a more robust transfer of these parameters within a force field, defined by its fundamental radii for intermolecular interactions. With this observation in mind, our objective in establishing the structural database for use in the parametrization of eq 3 was to maximize the diversity in the atomic environment (extent of burial) within a fixed “palette” of atom types (i.e., vdW radii). To sample a variety of buried environments, 22 globular proteins and 22 nucleic acid strands were selected and combined with small molecule atomic environments generated from the 20 amino acids, 210 dipeptides (generated by combining the amino acids in pairs), 5 nucleotides, and 15 dinucleotides (from pairs of nucleotides). These molecules offer sufficient variation in solvent-excluded volume to permit a variety of effective Born radii to be sampled and considered in the optimization of eq 3. The Protein Data Bank[24] codes for the 22 globular proteins and 22 nucleic acid strands, respectively, are as follows: 1ajj, 1bbl, 1bor, 1bpi, 1cbn, 1fca, 1frd, 1fxd, 1hpt, 1mbg, 1neq, 1ptq, 1r69, 1sh1, 1svr, 1tsg, 1uxc, 1vii, 1vjw, 2erl, 2pde, 451c, 2d47, 197d, 137d, 160d, 281d, 240d, 243d, 246d, 257d, 320d, 317d, 1d46, 1d56, 194d, 113d, 289d, 287d, 298d, 296d, 206d, 263d, and 271d.

Parametrization of the expression from eq 3 above can be considered to be dependent on the particular parameter set and topology representation one wishes to employ, e.g., all hydrogen atoms being represented explicitly, only polar hydrogen atoms, DNA versus proteins. Thus we developed two distinct “composite” databases reflecting our objective of deriving parameters for all hydrogen representations of CHARMM’s param22 protein[25] and nucleic acid[26] parameter sets as well as the polar hydrogen parameter set from param19.[27,28] In Table 1 we indicate the number of molecular species and the number of atomic environments represented by our database when partitioned as just noted.

Structures were prepared for the electrostatic analysis as follows. Single amino acids and nucleotides were built in an all trans conformation. All angles and bond lengths were taken from the appropriate parameter set as noted above. Dipeptides and dinucleotides were constructed using the default configuration from the appropriate CHARMM parameter set and subsequently minimized for 500 steps using the adopted basis Newton−Raphson algorithm. Minimizations were done in a vacuum (constant dielectric of 1.0) using a nonbonded cutoff value of 8 Å. Nonbonded interactions were truncated using a switching function between 6.5 Å and 7.5 Å. Proteins and nucleic acid strands were taken directly from the Brookhaven Protein Data Bank, and protons were added according to the appropriate parameter/topology set. No minimization was performed on the macromolecular structures. Radii and charges (for molecular solvation free energy calculations) were taken directly from the corresponding param19 and param22 CHARMM parameter sets. Following similar work by Schaefer and Karplus[12] as well as Still and co-workers,[10] we set the polar hydrogen radii in the param22 parameter set to 0.8 Å to eliminate pathologies that occur for very small intrinsic radii.

II.C. Finite Difference Poisson−Boltzmann Calculations. From eq 2 above we see that the generalized Born radius for an atom in a specific molecular environment is related to the Born polarization energy of an equivalent unit-charged ion in the dielectric boundary defined by the molecule containing atom *i*. Thus, to establish the reference generalized Born radii for atoms in our database(s), α_i, we calculated the polarization energy of each atomic site from the numerical solution to Poisson’s equation using a finite difference Poisson−Boltzmann method from DelPhi.[21] The solvation calculations were performed on each atom of every molecule in the two composite databases. The atom of focus was given a unit charge, and the charges of all other atoms in the corresponding molecule were set to zero. The van der Waals radii used in these calculations were obtained from the corresponding CHARMM parameter set, i.e., $R_{\min}$ values for atom types in the CHARMM parameter set were employed as atomic radii in defining the dielectric boundary. Solvation energies of complete molecules were computed for use in a second stage of optimization as discussed below. The van der Waals radii and charges, used in these DelPhi calculations, were both obtained from the corresponding CHARMM parameter set. A solvent probe radius of 1.4 Å was used in establishing the boundary between high dielectric exterior and low dielectric interior regions. The interior dielectric constant was set to 1, and an exterior dielectric constant of 80 was used to represent a high dielectric solvent like water. The spacing between grid points was chosen to be 0.25 Å, and each molecular system occupied 70% of the grid volume to avoid artifacts associated with being too close to the grid boundary. The large number of atoms contained within the param22 nucleic acid strand and protein databases required that we reduce the resolution of the grid to 0.5 Å for these calculations. However, testing the coarse grids revealed that this change did not significantly affect the final optimized parameters (data not shown).

II.D. Fitting Procedure. The objective of the fitting procedure is to obtain the optimal set of parameters for eq 3 to reproduce the Born radii computed from the Poisson equation and eq 2. The functional form associated with the fitting parameters, given in eq 3, has previously been demonstrated to reliably reproduce solvation energies[10] as well as partition coefficients[29] of small molecules in concert with the generalized Born equation (eq 1). However, it has been observed, by ourselves and others, that systematic errors in the generalized Born solvation energy occur in larger systems.[30] To eliminate this systematic error, we extend the fitting procedure used by

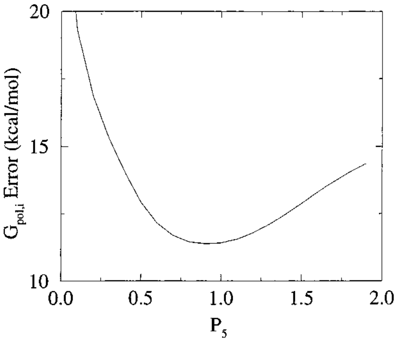

Figure 1. RMSD error in $G_{pol,i}$ as a function of P_5. P_1-P_4 and λ are reoptimized to provide the minimum error for each value of P_5.

Still through a second optimization of the van der Waals scaling parameter (λ) with respect to molecular solvation energies. Below we discuss the optimization of P_1-P_5 and λ based on the atomic Born radii, followed by details of the readjustment of λ to best fit molecular solvation free energies.

$G_{pol,i}$ Optimization. Using a linearized version of Still's empirical function for $G_{pol,i}$, we optimized the P_1-P_4 and λ parameters as a function of the only remaining nonlinear term, P_5. We did this by choosing discrete values of P_5 from 0.0 to 3.0 in steps of 0.1. The range and step size were determined to be adequate given the smoothness of the RMSD atomic solvation error as a function of P_5, shown in Figure 1. For each discrete value of P_5, P_1-P_4 and λ were optimized by multivariate linear regression. The complete parameter set corresponding to the minimum average RMSD between eq 3 and the PB values was selected as optimal.

Re-fitting of λ To Optimize Molecular Solvation Free Energies, G_{pol}. We found that although the fitting procedure described above achieved the optimal fit for Born radii, it did not give the optimal result for the total molecular solvation energy, and systematic errors occurred depending on the nature and composition of our database. To compensate for this systematic error in the molecular solvation energy, we also examined reoptimization of the van der Waals scaling parameter, λ.

This fitting procedure used a binary search algorithm[31] to minimize the unsigned error between G_{pol} obtained from the generalized Born equation (eq 1) and that obtained from the FDPB corrected reaction field. Born radii were computed using fixed P_1-P_5 values obtained from Born radii optimization described previously. The van der Waals scaling parameter was allowed to vary in order to compensate for the observed systematic error in the molecular solvation energy. A flowchart describing the fitting procedure is given in Figure 2.

II.E. Conformational Energies and Dynamics. We investigated the ability of the generalized Born model described above to reproduce relative solvation energies with respect to a molecular dynamics trajectory. Its success in reproducing energy changes relates directly to its value in determining forces for use in molecular dynamics. To examine this aspect of the generalized Born model we used conformations of two protein systems that were produced from unfolding trajectories in explicit solvent. The first was that of one of the helical segments of GCN4 leucine zipper [Brooks and Young, unpublished data], and the second was from extensive folding free energy calculations on a small helical protein, fragment B of staphylococcal protein A.[32]

To test the effectiveness of the Born model in replacing explicit solvent in molecular dynamics simulations, we also performed several 1 ns, room temperature simulations of the small α/β protein, segment B1 of streptococcal protein G (GB1).[33] We compared properties computed from the trajectories taken from simulations using uniform dielectric constants of $\epsilon = 1$ and $\epsilon = 80$, a distance-dependent dielectric with coefficient $\epsilon = 2.0$, as well as a simulation in explicit solvent. GB1 was chosen due to its size and consistency with systems used in the optimization of the Born equation and also because its native state dynamics in explicit solvent had been studied in this laboratory.[18] The starting structure for all dynamics runs corresponded to a relaxed form of the original crystal structure. Nonbonded cutoffs were not used during propagation of the dynamics of the "implicit" solvent calculations. However, we do consider examples in which cutoffs are used. In these cases, we employ strict cutoffs, i.e., no switching or shifting of the long-range interactions. As in previous calculations, the dielectric constant value used in the generalized Born model was set to 80 to represent a highly polarizable medium such as water.

In the protocol we employed for molecular dynamics, the structure was first relaxed in the presence of the "implicit" solvent using 1000 steps of energy minimization or until the energy change between subsequent steps was less than 0.001 kcal. Following the relaxation of the GB1 polypeptide, it was subjected to a total of 1 ns of molecular dynamics at a constant temperature of 298 K. The time step used was 2 fs and SHAKE was used to eliminate the high-frequency hydrogen/heavy atom bond vibrations. The last 600 ps were used in the analysis. This approach corresponds to that used in the explicit water simulations.[18]

III. Results and Discussion

III.A. Optimized Parameters. Parameters for the two composite databases are listed in Table 2. We note that a combined protein and nucleic acid structure database was used to fit the param22 "all hydrogen" force field. The results demonstrate some variability between the two force fields; however, most values for corresponding fitting parameters are reasonably close. This consistency lends credence to the physical model connected to the analytical approximation proposed by Still and co-workers (eqs 1 and 3).

We also performed a jackknife test in order to verify that the size of each composite database was sufficient to avoid overfitting. We removed 20% of the atoms (corresponding to eliminating randomly selected molecules from the databases) from the param19 and param22 molecular databases and refit $G_{pol,i}$, (eq 3). Root-mean-squared errors were calculated for the removed atoms using the jackknifed and original parameter sets. This jackknife procedure was carried out six times, using different sets of removed atoms, in order to estimate the standard deviation of the parameters and RMSD errors.

The parameters obtained from the reduced databases strongly resemble those obtained from the complete databases and demonstrate very small deviations from their average value (Table 2). In addition, the errors in atomic polarization energy calculated for the selected molecules using the two parameter sets were also very similar (Table 2). We conclude that the databases we have selected are diverse enough to avoid problems associated with overfitting.

As discussed previously, we were able to improve the fit to molecular solvation energy by reoptimizing the λ parameter. The λ parameter was re-fit within the context of the generalized Born equation against molecular solvation energies calculated with the Poisson equation. The newly optimized λ parameters were derived for each of the nine component databases, and the results are shown in Table 3. The components are separated based on the size of the molecules. We and others have observed

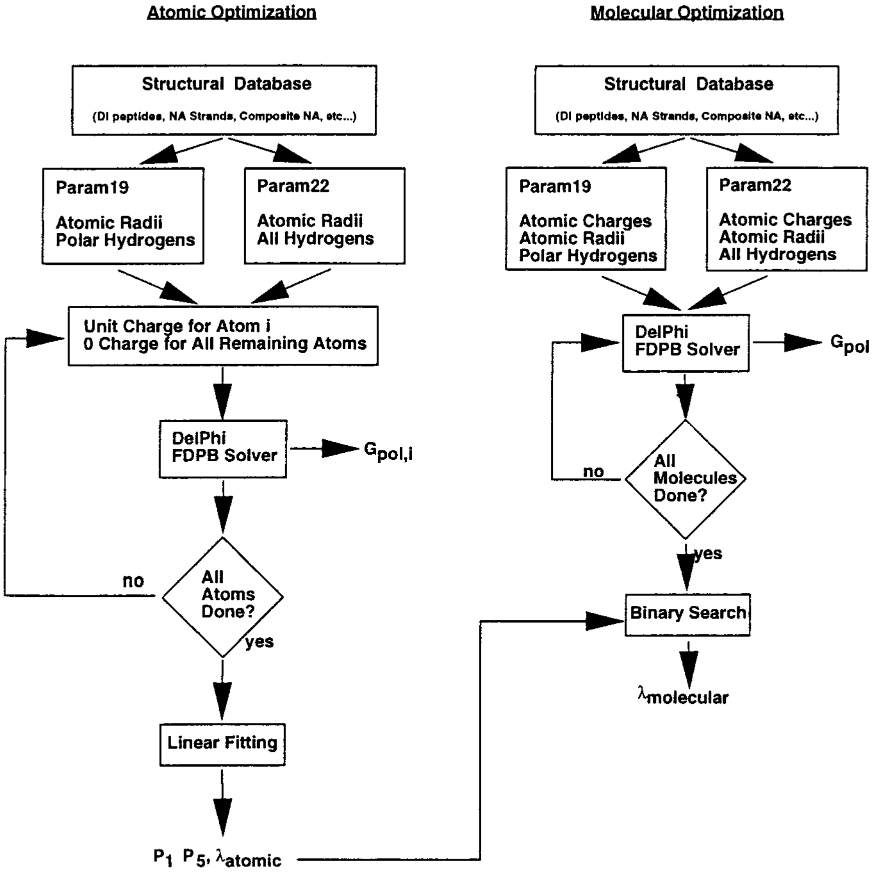

Figure 2. Flowchart describing the fitting procedure starting with optimization of the P_1-P_5 and λ atomic parameters for atomic solvation energies and leading into the reoptimization of λ for a more accurate representation of the molecular solvation energies.

TABLE 2: Parametrization of Eq 3 Using Composite Databases of Proteins and Nucleic Acids

parameters	param19	param22	param19 jackknife (std dev)	param22 jackknife (std dev)
P_1	0.415	0.448	0.412 (0.006)	0.449 (0.003)
P_2	0.239	0.173	0.248 (0.040)	0.172 (0.004)
P_3	1.756	0.013	1.651 (0.093)	0.013 (0.001)
P_4	10.51	9.015	10.48 (0.37)	8.92 (0.09)
P_5	1.1	0.9	1.1 (~0)	0.9 (~0)
λ_{atomic}	0.759	0.793	0.764 (0.007)	0.794 (0.003)
error (kcal/mol)	11.27	9.41	11.26 (0.221)	9.40 (0.08)
error20[a] (kcal/mol)	11.11/0.75	9.69/0.30	11.31 (0.91)	9.46 (0.31)

[a] The entry "error20" denotes the average RMSD error in $G_{pol,i}$ for the 20% of atoms removed.

TABLE 3: Re-fitting of the λ Parameter for Greater Precision in Calculating Molecular Solvation Energies

component (param19)	$\lambda_{molecular}$	component (param22)	$\lambda_{molecular}$
AA single	0.770	AA single	0.797
AA di	0.744	AA di	0.776
proteins	0.730	proteins	0.705
		NA single	0.770
		NA di	0.782
		NA strands	0.746

that size, in combination with a reasonable charge distribution, significantly influences the GB solvation energy relative to the FDPB value. Although this does not have a significant effect on the relative solvation energies or forces in a molecular dynamics simulation, it may be useful to use optimized λ parameters in more focused studies.

One common trend among the parameters is that the van der Waals scaling parameter, λ, is always less than 1. To achieve the best possible fit, whether for atomic or molecular solvation energies, atomic radii are consistently reduced. This affects a corresponding change in the Born radii, α. In addition, scaling parameters that have been re-fit to molecular solvation energies show a further decrease relative to scaling parameters, fit to atomic solvation energies. The result is a change in the self-polarization component of the generalized Born function toward more negative values. The cross-polarization term is also dependent on the scaling of atomic radii and, as predicted, becomes more positive. However, this change is not enough to compensate for the change in the self-energy.

We also note that there seems to be a consistent trend toward smaller values for the P_2-P_5 parameters optimized against explicit hydrogen structures (param22 composite database). These smaller parameters attenuate the dielectric shielding of a charge by surrounding atomic volumes. This attenuation could compensate for the larger number of atoms within the explicit hydrogen model which is most apparent in the larger number of angle terms associated with this model.

III.B Performance. The least-squares optimization of eq 3 using the composite databases listed above (Table 1) yields a reasonable set of parameters that can be used to compute solvation energies for a wide range of molecules (Table 2). Average unsigned errors were computed for each of the component databases using parameters optimized on the corresponding composite database as noted in Table 4.

The effectiveness of the reoptimization of the van der Waals scaling parameter with respect to molecular solvation energies is demonstrated in the right-most columns of Table 4. The new parameter, $\lambda_{molecular}$, significantly reduces the error for all

TABLE 4: Average Unsigned Error for Computed Solvation Energies of the Component Databases Using $\lambda_{\mathrm{atomic}}$ (Optimized To Reproduce Born Radii) and $\lambda_{\mathrm{molecular}}$ (Optimized To Reproduce Molecular Solvation Energies)[a]

database		% error ($\lambda_{\mathrm{atomic}}$)	% error ($\lambda_{\mathrm{molecular}}$)
single amino acids	(param19)	9.9	7.7
dipeptides	(param19)	9.5	5.6
proteins	(param19)	8.2	2.5
single amino acids	(param22)	5.3	5.7
dipeptides	(param22)	10.5	5.8
proteins	(param22)	24.1	8.4
single nucleotides	(param22)	11.7	2.3
dinucleotides	(param22)	3.6	0.8
nucleic acid strands	(param22)	2.4	0.8

[a] $$\text{error} = \frac{1}{N}\sum_{i=1}^{N}\frac{|GB_i - PB_i|}{PB_i}$$

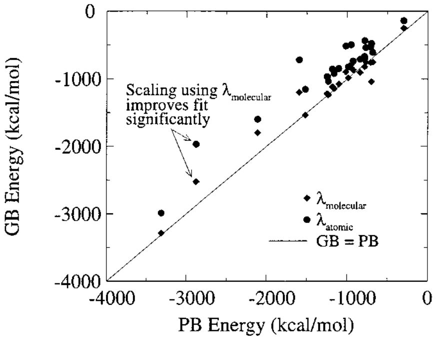

Figure 3. Example of the effect of reoptimizing the van der Waals scaling parameter, λ, with respect to molecular solvation energies for proteins with the param22 parameter set.

databases examined thus far. Specifically, reoptimization of this parameter effectively removes the effect of line shifting observed in the Still model as systems increase in size.[30] In addition, since the effect of this parameter is dependent on the composition of the molecule in terms of atomic radii, optimization of λ improves the correlation as well as the absolute error with respect to the FDPB results (Figure 3).

III.C. Conformational Energy Changes. Assessing the accuracy of the model on databases used in the optimization is one key to demonstrating the model's performance. To extend our understanding of the model, we examined two protein structures not contained within the training set. Specifically, we examined a range of conformations of protein A and the GCN4 single helix taken from unfolding trajectories performed in explicit solvent[32] [Brooks and Young, unpublished data]. Both of these systems were simulated in explicit solvent and used the CHARMM param19 polar hydrogen force field. The solvent was removed and the solvation energy calculated from generalized Born was compared to the corresponding FDPB results. The results are consistent with the error observed for proteins using the param19 parameters. Protein A and GCN4 conformations had errors of 8.9% and 7.1%, respectively.

We were most interested in how well the GB model reproduces changes in the solvation energy relative to FDPB results. To make such a comparison, the pairwise, conformationally dependent GB and PB solvation energy differences, ΔGB and ΔPB, respectively, were calculated, and the average error was computed. The use of each conformation as the reference point eliminates problems associated with choosing a single conformation whose error with respect to the PB result is not in line with the rest of the population. The ΔGB/ΔPB average percent unsigned error for protein A and GCN4 was 1.6% and 1.1%, respectively. The strong correlation between ΔGB and ΔPB is an indication that forces derived from this model can be used to drive a molecular dynamics simulation (Figure 4).

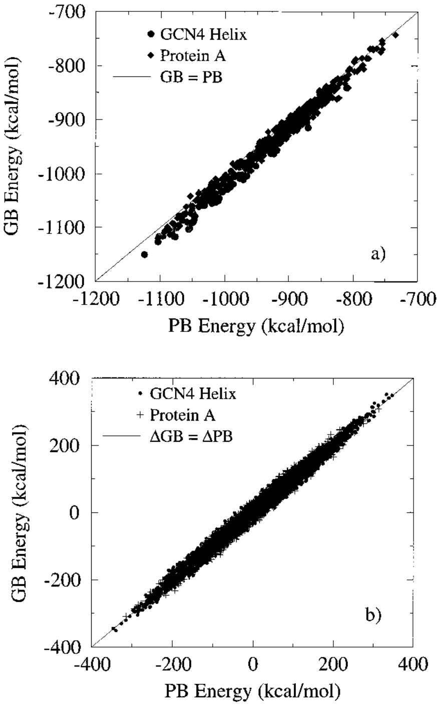

Figure 4. Solvation energies for multiple conformations of protein A and GCN4 single helix: (a) absolute solvation energies, PB vs GB, and (b) relative solvation energies, ΔPB vs ΔGB.

In addition to electrostatic solvation energy, however, it is possible that the hydrophobic component to the solvation energy will also yield significant forces that should be considered. To investigate this issue, we examined the GB electrostatic solvation energy and the hydrophobic solvation energy as a function of conformation for the unfolding of protein A and GCN4. The hydrophobic solvation energy was represented by a simple linear function of the surface area using a coefficient of 5 cal/mol/Å^2 as suggested by Sitkoff et al.[6] and a solvent probe radius of 1.4 Å. The results demonstrate that fluctuations in the electrostatic component of the solvation energy dominate those observed in the hydrophobic component. This suggests that, in terms of solvation, forces used in molecular dynamics will be dominated by the electrostatic component of the solvation free energy.

III.D. Dynamics of GB1 Using a Generalized Born Solvent. To directly test the effectiveness of the Born model in replacing explicit solvent in molecular dynamics simulations, we performed a 1 ns room temperature simulation of the protein GB1 employing the generalized Born model (eq 1). The full and correct implementation of eq 1 in dynamics requires that the derivative of the Born radius with respect to conformation be developed. This derivative was determined, incorporated into the CHARMM package, and used in conjunction with the electrostatic forces to drive the dynamics of the protein system. In the dynamics utilizing this solvent model, the Born radii and corresponding forces were updated at every time step, thereby ensuring the correct relationship between the energy function and its derivative. We also wished to explore the behavior of this implicit solvent model with others that might be employed. For this purpose, we compared properties from the simulation employing the generalized Born "water" model with trajectories taken from four other 1 ns simulations using different solvent

TABLE 5: RMSD Values (Å) Comparing Average Structures Obtained from 600 ps of Production Dynamics of a 1 ns Trajectory[a]

	GB 14 Å cutoff	GB	RDIE (eps = 2)	CDIE (eps = 1)	CDIE (eps = 80)	explicit
GB 14 Å cutoff						
GB	0.53					
RDIE (eps = 2)	2.03	2.03				
CDIE (eps = 1)	2.78	2.73	2.16			
CDIE (eps = 80)	2.59	2.43	3.12	4.10		
explicit	1.24	1.23	1.80	2.82	2.78	
crystal	1.09	1.27	2.13	2.99	2.81	1.50

[a] The generalized Born implicit solvation model appears to represent the original crystal structure better than any other solvation model above. GB parameters were taken from the optimization of the param19 composite database (Table 2).

models: uniform dielectric of $\epsilon = 1$; uniform dielectric of $\epsilon = 80$; distance-dependent dielectric with coefficient $\epsilon = 2.0$; explicit solvent.

Over the course of the 1 ns trajectory, the simulation utilizing the generalized Born solvent model maintained the secondary and tertiary structure of GB1. The average structure from the generalized Born simulation is most similar (as measured by backbone atom root-mean-square deviation, RMSD, of the model structure from other reference structures) to the crystal structure and the average structure from the explicit solvent simulation. The generalized Born model outperformed all other methods by these criteria. The RMSD similarity between average structures taken from all simulations is given in Table 5. In addition, the average structure from a simulation using a 14 Å cutoff and the generalized Born implicit solvent is compared. These results closely mimic those from the corresponding GB simulation using no cutoff. Details concerning the implementation of cutoffs within the GB formalism are discussed in the following section.

The time evolution of the RMSD relative to the crystal structure for these various simulations demonstrates that the generalized Born model stabilizes the structure near the compact crystal structure. The explicit solvent model shows very similar properties in terms of structural stability. The r-dielectric model appears to be relatively stable; however, it quickly finds a structural ensemble that is more distant from the crystal structure. The uniform dielectric models do not appear to be stable and are further from the crystal structure than any other solvent model. The figure below shows the time evolution of the RMSD from the crystal structure for the six simulations described above (Figure 5).

Finally, we examined the RMS fluctuations averaged over individual residues within the GB1 segment for each of the six simulations. The fluctuations observed in the explicit solvent and generalized Born model are very similar, while those observed for the other solvent models are larger. The only significant differences between the residue-averaged RMS fluctuations for the explicit solvent and generalized Born model occurred in flexible loop regions connecting secondary structure elements. A plot of the fluctuations as a function of residue for each of the six simulations is given in Figure 6.

III.E. Using Cutoffs. It is possible to speed up the single-point energy and dynamics calculations even further with the use of cutoffs. We have investigated their application using, as our test set, 22 proteins from the param19 database. Cutoffs reduce the $O(n^2)$ calculation to $O(n)$ without significantly affecting the single-point energies nor the direction and magnitude of the calculated forces.

As mentioned above, the set of structures we examined were the 22 proteins within the param19 database. The generalized Born solvation energy can be effectively split into two components: the self-polarization energy and the cross-polarization energy. The self-polarization energy depends only on the Born radii, and this depends only on cutoffs applied in the Born radii calculation (eq 3). The cross-polarization energy is dependent on the Born radius cutoff as well as the cutoff applied in the generalized Born equation (eq 1). Errors observed in the self-polarization energy for a typical 14 Å cutoff average around 1.6% while those observed for the cross-polarization energy average around 21%. From this observation, it is apparent that the generalized Born function will be more sensitive to errors than the Born radius or $G_{\mathrm{pol},i}$ calculation. This is consistent with the fact that $G_{\mathrm{pol},i}$ is proportional to $1/r^4$ while the generalized Born function is proportional to $1/r$. Average unsigned percent errors are given over a range of cutoffs for the 22 proteins, represented with the CHARMM param19 parameters in Table 6.

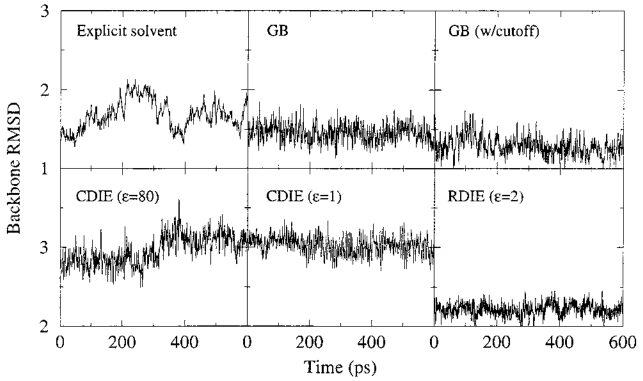

Figure 5. Time dependence of structural differences in dynamics simulations. Root-mean-square deviations (RMSD, in angstroms) from the crystallographic structure of protein G (1pgb) versus time from 1 ns simulations using various approximate solvent models.

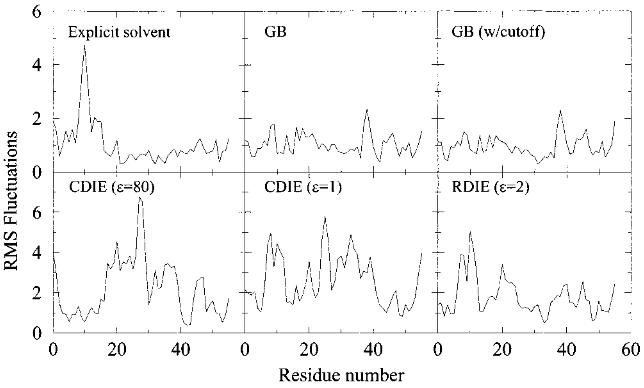

Figure 6. Residue-averaged RMS fluctuations for protein G from 1 ns simulations using various approximate solvent models.

Although the errors in the cross-polarization and the total GB energy indicate a problem in applying cutoffs to the generalized Born equation, we can take advantage of the well-known fact that the Coulomb interaction energy and the electrostatic solvation energy are strongly anti-correlated. Assuming our solvation model is mimicking aspects of real solvent, we should observe compensatory effects as a result of this anti-correlation. This appears to be the case.

When the behavior of the total electrostatic energy of the system is noted, very small errors are observed even for relatively short cutoffs. A 14 Å cutoff yields a 0.7% average error in the param19 protein database. Table 6 demonstrates the rapid convergence of the total electrostatic energy for these example stuctures.

TABLE 6: Errors in Generalized Born Solvation Energy Components[a]

cutoff (Å)	self-energy (%)	cross energy (%)	GB energy (%)	GB + Coulomb energy (%)
10	5.5	21.7	22.2	1.3
12	3.0	22.1	14.4	1.0
14	1.6	20.9	15.2	0.7
16	0.8	17.7	14.6	0.3
18	0.4	12.7	10.1	0.2
20	0.2	10.4	8.5	0.1

[a] Small relative errors in the self-energy indicates that the Born radii will not be significantly affected by applying cutoffs. Compensation between the Coulomb and GB terms significantly reduces the overall error incurred due to cutoffs (shown in last column).

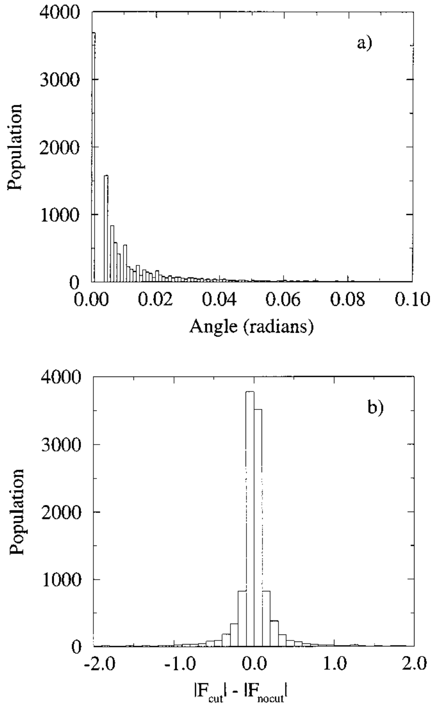

Figure 7. Force errors from interaction cutoffs in generalized Born calculations: (a) distribution of the angle between force vectors calculated using a 14 Å cutoff and an infinite cutoff; (b) distribution of the corresponding magnitudes of the forces.

In addition to analyzing the effects of cutoffs on energy calculations, we also examined the effects on the atomic forces. To explore these consequences, we considered both the angle between and the differences in the magnitude of the electrostatic forces employing the generalized Born solvent model with and without cutoffs. The angles between force vectors were calculated using a 14 Å cutoff and an infinite cutoff for each atom in the param19 protein database (11995 atoms). The vast majority of angles were very close to 0°, indicating that the force vectors are parallel. Similar results were obtained by examining the change in the force magnitudes. Distributions of the angle between force vectors and their magnitude differences are given in Figure 7. These results indicate that reasonable cutoffs do not significantly affect the energetics or the dynamics of systems using the generalized Born implicit solvent model.

We also ran a room temperature simulation of the B1 segment of protein G using the GB implicit solvent and a cutoff of 14 Å. The average structure resulting from the 600 ps of production dynamics very closely resembles that obtained from the GB simulation using no cutoffs (Table 5). In addition, the average structure taken from the simulation using cutoffs was similar to both the crystal structure and the average structure resulting from the explicit solvent simulation (Table 5). This example supports the validity of cutoffs in the context of the GB model.

TABLE 7: Computation Time for 1000 Steps of Dynamics Using the B1 Segment of Protein G as the Test System[a]

solvent model	time (min)	relative time
vacuum ($\epsilon = 1$)	1.24	1*X*
CDIE ($\epsilon = 80$)	1.24	1*X*
RDIE ($\epsilon = 2$)	1.11	0.9*X*
GB	5.88	5*X*
explicit solvent	220	177*X*

[a] All simulations used a cutoff.

It is also of interest to examine the computation time required by common continuum models and explicit solvent. For the purposes of timing, all models utilized a cutoff to be consistent with commonly used simulation conditions. Table 7 lists the time required to run 1000 steps of dynamics with protein G for each of the solvation models on a 180 MHz R10000 processor within an SGI Origin 200 workstation. It also lists the relative times compared to a vacuum simulation. The GB model is approximately five times slower than the vacuum simulation, but over 35 times faster than the explicit solvent model using similar cutoff conditions. In addition, it is interesting to note that the solution of Poisson's equation for a protein of this size on the same processor using the DelPhi program takes about 84 s/conformation, or approximately 240 times slower than the time it takes for a single configuration using the GB model.

We have demonstrated in this section that it is possible to maintain the reliability of the GB model using cutoffs. Insignificant errors in the self-polarization energy term due to the r^{-4} dependence as well as compensation between the cross-polarization and Coulomb energy terms make cutoffs a reasonable approach to reducing computation time. In addition, we have shown that the generalized Born model in conjunction with cutoffs provides a very rapid method for including important solvent effects.

IV. Conclusions

The objective of this work was to develop a parametrization of the generalized Born model of Still and co-workers that was consistent with a given force field and capable of reproducing the electrostatic solvation free energy as calculated by the finite difference solution to the Poisson equation (FDPB). A highly relevant class of molecules, namely polypeptides and nucleic acid structures, has been used as training sets for optimizing a linearized form of the analytical approximation for the Born radius introduced by Still. The parametrization from these structural databases was highly effective in reproducing the solvation free energy determined by the finite difference solution to the Poisson–Boltzmann equation. Further, optimizing with composite databases is able to provide parameter sets that are applicable to a wide range of biomolecules.

The use of the van der Waals scaling parameter, λ, significantly increases the accuracy of the fit for both atomic and molecular solvation energies. Specifically, reoptimization of this scaling parameter effectively removes the effect of line shifting observed as systems increase in size. In addition, since the effect of this parameter is dependent on the composition of the molecule in terms of atomic radii, optimization of λ improves the correlation as well as the absolute error with respect to the FDPB results.

Studies of the relative changes in solvation free energy with conformation using protein A and a single helix of GCN4 suggest the high fidelity of this method in reproducing forces derived from the FDPB solvation energies. In addition, the dynamics study of the B1 segment of protein G demonstrates the ability of the model to effectively replace explicit solvent. This is shown by the stability of the generalized Born simulation as well as the strong structural similarity of the average structure to both the crystal structure and the average structure from the explicit solvent simulation.

The generalized Born model from Still and co-workers, fit using amino acid and nucleotide structural databases, is capable of replacing solvent in cases where a mean field is an appropriate approximation. The increased speed of the generalized Born calculation relative to both the FDPB calculation and pairwise calculations of explicit solvent interactions suggests this approach as a viable alternative in the field of structural genomics where large numbers of calculations are required. The addition of solvent effects could improve molecular models, which are typically not reliable at predicting the large number of specific contacts required for biochemical resolution. Furthermore, applications in assessing ligand binding affinity for high-throughput screening calculations and in exploratory simulations of folding and dynamics are potentially interesting uses of such fast analytic implicit solvent models.

Acknowledgment. This work was supported through grants from the NIH (GM37554, GM48807, RR12255) and Novartis Pharmaceuticals. Their support is appreciated. In addition, discussions with Professor D. Case and his group members provided pivotal direction to this work. Their interest and input is greatly appreciated. Finally, support for the initial stages of this project from Molecular Simulations Inc. is acknowledged.

References and Notes

(1) Duan, Y.; Wang, L.; Kollman, P. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9897−9902.

(2) Daggett, V.; Kollman, P. A.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 285−304.

(3) Smart, J. L.; Marrone, T. J.; McCammon, J. A. *J. Comput. Chem* **1997**, *18*, 1750−1759.

(4) Gilson, M. K.; McCammon, J. A.; Madura, J. D. *J. Comput. Chem.* **1995**, *16*, 1081−1095.

(5) Brooks, C. L., III; Case, D. A. *Chem. Rev.* **1993**, *93*, 2487−2502.

(6) Sitkoff, D.; K. A., S.; Honig, B. *J. Am. Chem. Soc.* **1994**, *98*, 1978−1988.

(7) Lazaridis, T.; Karplus, M. *Science* **1997**, *278*, 1928−1931.

(8) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.

(9) Born, M. Z. *Phys.* **1920**, *1*, 45.

(10) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.

(11) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(12) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.

(13) Klamt, A.; Schuurmann, G. *Chem. Soc. Perkin Trans.* **1993**, *2*, 799−805.

(14) Srinivasan, J.; Cheatham, T. E., III; Piotr, C.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.

(15) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(16) Cammi, R.; Tomasi, J. *J. Chem. Phys.* **1994**, *100*, 7495−7502.

(17) Warwicker, J.; Watson, H. C. *J. Mol. Bio.* **1982**, *157*, 671−679.

(18) Sheinerman, F. B.; Brooks, C. L., III *Proteins* **1997**, *29*, 193−202.

(19) Sheinerman, F. B.; Brooks, C. L., III *J. Mol. Bio.* **1998**, *278*, 439−456.

(20) Sheinerman, F. B.; Brooks, C. L., III *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1562−1567.

(21) Gilson, M. K.; Honig, B. *Proteins* **1988**, *4*, 7−18.

(22) Gilson, M. K.; Honig, B. *J. Comput. Aid. Mol. Des.* **1991**, *5*, 5−20.

(23) Kirkpatrick, S.; Gelatt, C.; Vecchi, M. *Science* **1983**, *220*, 671−680.

(24) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. *J. Mol. Bio.* **1977**, *112*, 535−542.

(25) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Gao, 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., III; 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*, 3586−3616.

(26) MacKerell, A. D., Jr.; Wiorkiewicz-Kuczera, J.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946−11975.

(27) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902−1921.

(28) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(29) Best, S. A.; Merz, K. M. J.; Reynolds, C. H. *J. Phys. Chem. B* **1997**, *101*, 10479−10487.

(30) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem. B* **1997**, *101*, 1190−1197.

(31) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*, 2nd ed.; Press Syndicate of the University of Cambridge: New York, 1992.

(32) Boczko, E. M.; Brooks, C. L., III *Science* **1995**, *269*, 393−396.

(33) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721−4729.