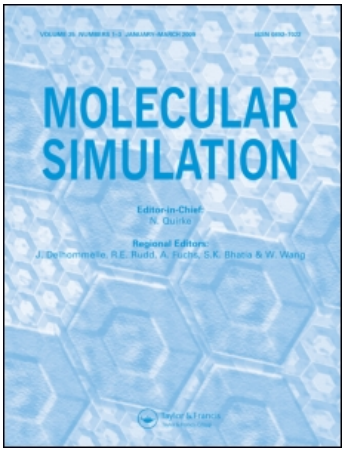

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Parameterization and Application of an Implicit Solvent Model for Macromolecules

Brian N. Dominy[a]
[a] Department of Molecular Biology, TPC6, The Scripps Research Institute, La Jolla, California

To cite this Article Dominy, Brian N.(2000) 'Parameterization and Application of an Implicit Solvent Model for Macromolecules', Molecular Simulation, 24: 4, 259 — 274
To link to this Article: DOI: 10.1080/08927020008022375
URL: http://dx.doi.org/10.1080/08927020008022375

Molecular Simulation, 2000, Vol. 24, pp. 259–274
Reprints available directly from the publisher
Photocopying permitted by license only

PARAMETERIZATION AND APPLICATION OF AN IMPLICIT SOLVENT MODEL FOR MACROMOLECULES

BRIAN N. DOMINY

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

(Received April 1999; accepted May 1999)

As the field of theoretical biophysics begins to recognize systems of longer timescales and larger magnitude, rapid approaches for investigating these systems are required. One promising simplification of the typical system of a solute surrounded by water is the use of implicit solvation models. The generalized Born implicit solvent offers a rapid approach for computing the electrostatic effects of bulk solvent without the explicit representation of water molecules. This report describes the parameterization of a generalized Born (GB) model for protein and nucleic acid structures. As a demonstration of the usefulness of this approach, the GB model is applied toward the discrimination of misfolded and properly folded protein structures. This study attempts to illustrate the potential of the GB model for molecular dynamics simulations over longer timescales as well as for screening large structural databases.

Keywords: Generalized Born; implicit solvation; electrostatics; protein; nucleic acid; misfolded protein structures

INTRODUCTION

In order for molecular simulations to reveal useful information about the physical world, it is important to use accurate models to represent the system of interest. Equally important is the ability to simulate systems over time scales relevant to the processes of interest. Unfortunately, it is often the case that these two conditions are subject to an inverse relationship in which model accuracy must be sacrificed for speed or *vice versa*.

A common example of this competition between speed and accuracy is the treatment of bulk solvent [1, 2]. The inclusion of explicit solvent

molecules surrounding the solute of interest can increase the number of atoms in the system by an order of magnitude. Given this size increase, and the two-body nature of most empirical force fields, the inclusion of explicit solvent can increase the computational cost by two orders of magnitude. In studies where the interest is focused primarily on the solute and not on the solvent behavior, it would be advantageous to remove the water molecules from the simulation. However, simply excluding water from a simulation can have a negative impact on the structural ensemble of the solute and therefore the thermodynamic and kinetic processes being studied [3–5]. A compromise between the complete atomic representation of solvent and a lack of all solvent effects is required.

A useful approach for avoiding the computational cost associated with explicit water molecules is to assume that bulk solvent behaves to first order as a mean field [2]. In this approximation, the configuration of the solute determines an average interaction between the solute and a solvent field. This average interaction can be implemented through an additional potential energy term (or modification of an existing term), which mediates interactions between solute atoms as if an explicit solvent were present. Although the effect of the bulk solvent is represented, individual solvent molecules are not and as a result the computational cost associated with the system is substantially reduced.

The generalized Born (GB) model is one method for representing the electrostatic effects of bulk solvent as a mean field [6]. The generalized Born method, an extension of the Born model for calculating ionic solvation energies [7], computes a new interaction energy between two point charges within the solute. The equation is similar in form to a Coulomb potential, but of opposite sign to represent the effects of electrostatic shielding due to polar solvents (Eq. (1)). The magnitude of the generalized Born interaction, unlike a simple Coulomb potential, is mediated by the distance between a given charge and the bulk solvent. This parameter is called the Born radius (α).

$$G_{\text{pol}} = -166 \cdot \left(1 - \frac{1}{\varepsilon_{\text{water}}}\right) \sum_{i}^{N} \sum_{j}^{N} \frac{q_i q_j}{\sqrt{\left(r_{ij}^2 + \alpha_i \alpha_j e^{-D_{ij}}\right)}} \quad \text{with } D_{ij} = \frac{r_{ij}^2}{4\alpha_i \alpha_j} \tag{1}$$

Since the Born radius must be computed for each atom in a given solute conformation, it is important that a rapid method for the calculation be available in order to maintain the advantage of speed. We have adapted an analytical method for the calculation of the Born radii first implemented

by Still and co-workers for small solute molecules. The modifications made in this equation make practical the use of the generalized Born equation with macromolecular structures including proteins and nucleic acid strands. The method appears to be robust in reproducing electrostatic solvation energies obtained from the finite difference solution of the Poisson equation [8, 9]. Further, an analytical solution to the first derivative of the generalized Born energy makes possible the incorporation of this implicit solvent model into molecular dynamics simulations of macromolecules. Simulations performed using the generalized Born solvent appear to reproduce backbone root mean squared deviations (rmsd) and amino acid sidechain fluctuations as seen in explicit water simulations. In total, the generalized Born model appears to be efficient in replacing explicit solvent in both static and dynamic contexts.

In this work, we also describe an application of the generalized Born model in conjunction with the standard CHARMM force field [10] for the discrimination of properly folded and misfolded protein structures. Methods able to quickly screen protein conformations and accurately identify stable and unstable folds have great potential in the evolving fields of protein folding and structural genomics. One example of their application would be in conjunction with lattice folding methods for predicting the correct fold of a given protein sequence. Lattice methods using knowledge-based potentials have the capability of generating large numbers of potential protein folds [11 – 13]. Rapid screening functions, which may include an implicit solvation term, may be useful in further discriminating between useful and misleading conformations.

Misfolded structures are obtained from the EMBL and have been generated by threading sequences into protein structures of the same sequence length, but different global fold [14]. These compact, misfolded structures are locally minimized and distinguished using the CHARMM/GB force field. It is found that electrostatic interactions, both intramolecular and solvent mediated, can be used very effectively to identify the misfolded structures. We further find that misfolded structures are more soluble than their properly folded counterparts. This principle is shown to have potential implications for protein folding when a compact intermediate state is formed.

METHODS

The content of this paper can be logically divided into two sections. The first section describes the development and testing of a generalized Born

parameterization for proteins and nucleic acids. The second section describes an application of this model toward the discrimination of properly folded and misfolded protein structures. The methods section will be organized according to these divisions.

Implementation of the Generalized Born Model

The Poisson equation provides an exact solution for the electrostatic free energy of a charge distribution surrounded by a dielectric medium [15]. Since the Poisson equation can be solved analytically only for simple geometries, numerical solutions, such as the finite difference approach, have been applied to arbitrarily shaped charge distributions such as proteins [8]. One of the bottlenecks of many numerical solutions to the Poisson equation is that a solvent accessible boundary must be constructed for each solute configuration. This boundary would be used equivalently in the generalized Born equation, an approximation of the Poisson equation, for computing the burial depth (or Born radius) of each charge within the protein. In order to avoid the computational expense required for the construction of the solvent accessible boundary, Still and co-workers have developed a rapid, pairwise approach to estimating Born radii [16]. The original equation has been linearized so that a multivariate linear regression approach may be applied to fit the equation for the large number of atoms present in macromolecular structures [17]. A van der Waals radius scaling factor (λ) has been included in the equation in order to avoid systematic fitting errors observed in the original paper. The use of the generalized Born equation in conjunction with the empirically fit expression for the Born radii (Eq. (2)) reproduces the Poisson solution at a significantly reduced computational cost.

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

$$\text{where } CCF = \begin{cases} 1.0; & \left(\dfrac{r_{ij}}{R_{\mathrm{vdW},i} + R_{\mathrm{vdW},j}}\right)^2 > \dfrac{1}{P_5} \\ \left\{0.5\left[1.0 - \cos\left(\left(\dfrac{r_{ij}}{R_{\mathrm{vdW},i} + R_{\mathrm{vdW},j}}\right)^2 \cdot P_5 \pi\right)\right]\right\}^2; & \left(\dfrac{r_{ij}}{R_{\mathrm{vdW},i} + R_{\mathrm{vdW},j}}\right)^2 \leq \dfrac{1}{P_5} \end{cases} \quad (2)$$

The terms R_{vdW}, V and r_{ij} represent the van der Waals radius and volume of a given atom as well as the distance between atoms i and j respectively. The parameters $P_1 - P_4$ and $1/\lambda$ are fit by multivariate linear regression to a collection of atomic solvation energies. These values were obtained from the Poisson solution of a unary charged atom in the presence of an otherwise

uncharged molecule using the DelPhi application [9, 18]. Each atom in the training set of molecules is given a unit charge while all other atoms in the given molecule are left uncharged. The Poisson solution for such a system is the interaction of that ion with the solvent dielectric in the context of the rest of the molecule. The P_5 parameter, although non-linear, is restricted to a small range of values and is fit by computing the error over this range of P_5 given the previously fit values of $P_1 - P_4$ and $1/\lambda$. The P_5 value giving the minimum error is chosen as the optimal value and completes the parameter set. The error function used is the root mean squared error between the $G_{\text{pol},i}$ value and the Poisson solution over each atom in the training set. All Poisson equation calculations referred to in this study are accomplished using the finite difference solver implemented in DelPhi [9, 18]. Grid sizes are chosen so that the solute occupies 80 percent of the total grid volume. A grid spacing of 0.5 Å is used to compute atomic solvation energies while a 0.25 Å grid spacing is used to compute molecular solvation energies described below.

The training set includes all atoms from a collection of molecules including proteins and nucleic acid strands derived from CHARMM's polar hydrogen [10] and explicit hydrogen [19, 20] force fields (Tab. I). The atomic polarization energies from Eq. (2) are computed based on the intrinsic Born ionic solvation energy of the charged atom as well as the pairwise distances from the charged atom to all other atoms in the molecule. The volume and distance of each surrounding atom reduces the ion's solvation energy due to the displacement of the solvent dielectric. The Born radius is inversely related to the atomic polarization energy (Eq. (2)) and is therefore trivially obtained for use in the generalized Born equation (Eq. (1)).

TABLE I Training set of molecules used in the parameterization of the analytical expression for $G_{\text{pol},i}$. CHARMM's polar hydrogen model includes amino acid structures while the explicit hydrogen model includes both amino acid and nucleic acid structures

Class	*# Molec.*	*# Atoms*
Polar hydrogen force field (CHARMM Param19)		
Single AA	20	203
Di-AA	210	4263
Proteins	22	11995
Explicit hydrogen force field (CHARMM Param22)		
Single AA	20	324
Di-AA	210	6804
Proteins	22	19417
Single NA	5	151
Di-NA	15	921
Strands	22	13496

Following optimization of the $G_{pol,i}$ equation (or the Born radii), it has been demonstrated that a systematic error may still be present in the solution of the generalized Born equation. It is possible to re-optimize the λ parameter in Eq. (2) with respect to molecular solvation energies instead of atomic solvation energies in order to eliminate this systematic error. A binary search algorithm was implemented to re-optimize the λ parameter to fit the solvation energies of the molecules used in the training set [17]. The $P_1 - P_5$ parameters were held fixed at their previously fit values. The different values of lambda are termed λ_{atm} and λ_{mol} to distinguish the parameter optimized to fit $G_{pol,i}$ (Eq. (2)) from that fit to the generalized Born equation (Eq. (1)) respectively.

Finally, an analytical first derivative of the generalized Born equation was calculated and implemented into the CHARMM molecular mechanics package [10]. Using the generalized Born equation and its first derivative, it is possible to compute not only solvation energies for static solute conformations but also the forces required for molecular dynamics simulations.

Fragment B1 of protein G was chosen in order to test the ability of a generalized Born term to replace explicit solvent in a dynamic simulation. The structure was first minimized in the presence of the "implicit" solvent for 1000 steps or until the energy change between subsequent steps was less than 0.001 kcal. The protein was then subjected to 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 [21]. The first 400 ps were considered equilibration while the last 600 ps were used in the analysis. This approach corresponds to that used in the explicit water simulations of protein G previously executed in the lab [22–24].

Determination of Misfolded Proteins

In this section we discuss the application of the CHARMM/GB force field for the discrimination of properly folded and misfolded protein structures. First we describe the EMBL and PDB [25] structures and how they were refined for the CHARMM/GB calculations. We then detail the energy terms used in computing the static energy of the misfolded and correctly folded protein structures.

Structures from the EMBL databases of misfolded proteins [14] as well as the corresponding PDB structures were first re-optimized for analysis in the CHARMM/GB force field. Side chain conformations were optimized in the original work using a systematic search of rotomer conformations

followed by a short minimization in the GROMOS force field [26]. We re-optimized the structures by minimizing under successively reduced harmonic restraints within the CHARMM param19 force field [10]. The average atomic root-mean-square deviation (rmsd) following this relaxation was approximately 0.3 Å. The harmonic restraints were reduced from 30 kcal/mol Å to 5 kcal/mol Å in increments of 5 kcal/mol Å. At each level of harmonic restraint, structures were minimized until the energy change between two sequential steps was less than 0.001 kl. A switching potential between 6.5 Å and 7.5 Å was used to reduce computational time. The minimizations were run under vacuum conditions. A total of 26 EMBL misfolded proteins and the corresponding 23 PDB structures subjected to this re-optimization make up the test set used throughout this study. The PDB accession codes as well as the corresponding EMBL misfolded structure names are listed (Tab. VII).

We note that it is improbable that unrelated topologies could support native disulfide linkages, large heme groups, or other ligands without exhibiting an extremely unfavorable energy. Ligands and disulfide linkages were not considered in either misfolded or correctly folded structures in order to remove any bias in the stability that might trivially favor the native fold. This is consistent with previous studies of the EMBL database [14, 27].

Nonbonded energy components were analyzed in order to determine the overall lowest energy conformer of each folded/misfolded pair or triplet. The lowest overall energy is predicted to be the correctly folded structure. Intramolecular energy components included van der Waals and Coulombic interactions between nonbonded atoms. Solvent was accounted for using a generalized Born model for electrostatic effects and a term proportional to surface area for hydrophobic effects. The proportionality constant used in the hydrophobic term is 5 cal/mol Å^2 as taken from the work of Sitkoff *et al.* [28]. In order to include all long-range interactions, no cutoffs were employed in these energy calculations. A dielectric constant of $\varepsilon = 1$ was used to represent the protein interior in both the Coulomb and generalized Born calculations. A simple (unweighted) linear sum of these energy terms is used to discriminate between misfolded and properly folded proteins.

RESULTS AND DISCUSSION

Parameterization and Testing of Generalized Born

The generalized Born equation, with Born radii parameterized according to the procedure outlined in the previous section, is able to accurately

reproduce energies and forces as obtained from the Poisson equation. Optimized parameters were generated for two structural databases consisting of amino acid and nucleic acid structures (Tab. II). When the lambda parameter associated with the simple optimization of atomic polarization energies is used (λ_{atm}) the error between the generalized Born and Poisson solutions is generally 12% or lower (Tab. IV). The exception being the proteins from the explicit hydrogen force field (param22). All errors are reduced to less than 10% by using the λ_{mol} parameter optimized to reproduce molecular solvation energies (Tab. III). The re-optimization of the lambda parameter eliminates a consistent error observed previously in the literature for macromolecular structures [29].

Using the first derivative of the generalized Born equation including the analytical expression for the Born radius (Eq. (2)), it is possible to calculate forces consistent with this implicit solvent model. Molecular dynamics simulations were performed over the course of 1 ns on the GB1 segment of protein G. Five independent simulations compared the ability of various dielectric models in reproducing the properties observed in an explicit water simulation. These dielectric models include the generalized Born

TABLE II Optimized parameters for Eq. (2) obtained by fitting by multivariate linear regression to Poisson solutions. The param19 set contain amino acid based molecules with polar hydrogens and extended carbons representing aliphatic hydrogen atoms. The param22 set contain amino acid and nucleic acid based structures containing polar and aliphatic hydrogen atoms

Parameters	*Param*19	*Param*22
P_1	0.415	0.441
P_2	0.239	0.185
P_3	1.756	0.0124
P_4	10.51	9.60
P_5	1.1	0.9
λ_{atomic}	0.759	0.798
Error (kcal/mol)	11.27	9.41

TABLE III Lambda parameters optimized to reproduce molecular solvation energies of each class of molecule used in the training set. Molecular solvation energies are given by the Poisson solution for the fully charged molecules

*Param*19	λ_{mol}	*Param*22	λ_{mol}
AA Single	0.7703	AA Single	0.7957
AA Di	0.7437	AA Di	0.7743
Proteins	0.7295	Proteins	0.6941
–	–	NA Single	0.7684
–	–	NA Di	0.7793
–	–	NA Strands	0.7361

TABLE IV Errors in the generalized Born molecular solvation energies relative to the corresponding Poisson solution. Lambda parameters were optimized to fit atomic and molecular solvation energies

Molecule class		*% Error* (λ_{atm})	*% Error* (λ_{mol})
Single amino acids	(param19)	9.9	7.7
Dipeptides	(param19)	9.5	5.6
Proteins	(param19)	8.2	2.5
Single amino acids	(param22)	6.2	5.8
Dipeptides	(param22)	13.5	6.2
Proteins	(param22)	28.3	9.2
Single nucleotides	(param22)	15.4	2.3
Dinucleotides	(param22)	6.3	0.9
Nucleic acid strands	(param22)	3.1	0.9

TABLE V Root mean squared differences between average structures generated from 1 ns molelar dynamics simulations of fragment B1 of protein G. Also compares average structures to the original crystal structure

	GB 14 Å cutoff	*GB*	*RDIE (eps = 2)*	*CDIE (eps = 1)*	*CDIE (eps = 80)*	*Explicit*
GB 14 Å Cutoff	–	–	–	–	–	–
GB	0.5	–	–	–	–	–
RDIE (eps = 2)	2.0	2.0	–	–	–	–
CDIE (eps = 1)	2.8	2.7	2.2	–	–	–
CDIE (eps = 80)	2.6	2.4	3.1	4.1	–	–
Explicit	1.2	1.2	1.8	2.8	2.8	–
Crystal	1.1	1.3	2.1	3.0	2.8	1.5

model, a distance dependent dielectric with coefficient 2.0, and two continuum dielectric models using $\varepsilon = 1$ and 80.

In comparing the root mean squared distance (rmsd) between the average structures generated from these various simulations and the original crystal structure we find that the generalized Born and explicit solvent models remain similar to the original crystal structure (Tab. V). The average structures obtained from the GB and explicit solvent models are also quite similar to each other. In addition, the rms fluctuations of individual residues during the course of the simulations further illustrate that the explicit solvent and GB models are similar and also most stable (Fig. 1). Sequence regions demonstrating higher flexibility in these models correlate with loop structures in the protein, which are expected to have greater mobility.

Finally, we investigated the utility of cutoffs in the generalized Born equation. The advantage to using such methods is to simplify the quadratic $O(n^2)$ energy calculation to a linear $O(n)$ problem. The problem with such methods comes from truncating potentials that are significant over long distances such as the Coulomb or generalized Born terms, which dissipate

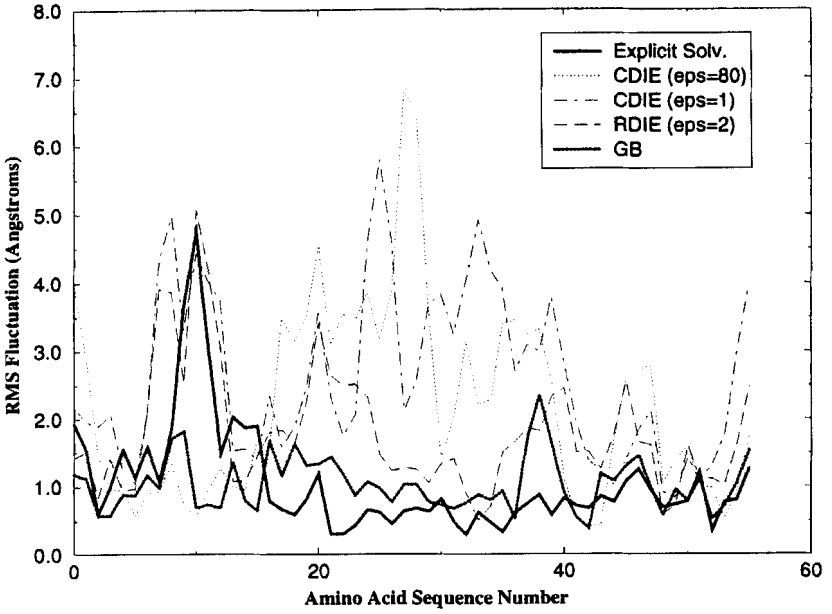

FIGURE 1 Sidechain fluctuations observed in the molecular dynamics simulation using various solvation models. Average RMS fluctuations are projected on thet *ordinate* while the primary sequence number is shown on the *abscissa*. Generalized Born and explicit solvent both show relatively low fluctuations. Regions of large fluctuations observed in GB and explicit solvent simulations correspond to loop regions of the protein G structure.

TABLE VI Percent errors between energy values computed at the given cutoff in relation to the value computed with no cutoff

Cutoff	*Self energy*	*Cross energy*	*GB energy*	*GB + Coulomb energy*
10 Å	5.5%	21.7%	22.2%	1.3%
12 Å	3.0%	22.1%	14.4%	1.0%
14 Å	1.6%	20.9%	15.2%	0.7%
16 Å	0.8%	17.7%	14.6%	0.3%
18 Å	0.4%	12.7%	10.1%	0.2%
20 Å	0.2%	10.4%	8.5%	0.1%

only as $1/r_{ij}$. Although the use of cutoffs on either term alone can produce errors, the sum of the Coulomb and generalized Born terms tend to produce a constant electrostatic potential within a relatively short range. This is due to the anti-correlated behavior well known with regard to the Coulomb and electrostatic solvation free energy. In the case of protein G a 14 Å cutoff can produce a 15% error in the generalized Born term alone relative to no cutoff, however only a 0.7% error in the sum of the Coulomb and GB terms (Tab. VI).

TABLE VII List of the PDB and EMBL structures and CHARMM/GB energy terms used in the discrimination of misfolded and properly folded protein structures. Values are given as the difference between the misfolded and properly folded energies (Properly folded – Misfolded) normalized for the number of residues in each sequence. All energy values are generally negative (*i.e.*, favoring the folded state) while the generalized Born term is generally positive (*i.e.*, favoring the misfolded state). This is due to sub-optimal charge pairing in the misfolded state

Name	(*misfold*)	$\Delta GB/Res$	$\Delta Coulomb/Res$	$\Delta NPSolv/Res$	$\Delta vdW/Res$	*Total* (*pdb*)	(*misfold*)	$\Delta Total/Res$
1bp2	1bp2on2paz	5.28	−7.49	**0.00**	−0.53	−7942.00	−7604.84	−2.74
1cbh	1cbhon1ppt	0.62	−0.11	−0.08	−1.08	−1608.95	−1586.08	−0.64
1fdx	1fdxon5rxn	**−2.10**	**2.73**	−0.06	−0.35	**−2728.95**	**−2741.00**	**0.22**
1hip	1hipon2b5c	5.49	−6.59	−0.05	−0.83	−4997.93	−4829.53	−1.98
1lh1	1lh1on2i1b	0.81	−2.89	−0.03	−0.27	−9663.51	−9299.77	−2.38
1p2p	1p2pon1rn3	2.58	−3.50	−0.02	**0.00**	−7714.33	−7596.76	−0.95
1ppt	1ppton1cbh	**−3.90**	**0.87**	**0.04**	**0.03**	−2078.76	−1972.70	−2.95
1rei	1reion5pad	1.35	−2.22	−0.03	−2.93	−11669.67	−10851.09	−3.83
1rhd	1rhdon2cyp	1.94	−3.31	**0.00**	−0.75	−17635.77	−17017.91	−2.11
1rn3	1rn3onlp2p	1.12	−1.82	−0.01	−0.66	−7597.77	−7427.96	−1.37
1sn3	1sn3on2ci2	4.99	−6.31	−0.04	−0.63	−4159.15	−4029.41	−2.00
	1sn3on2cro	0.86	−2.48	−0.01	−1.10	−4159.15	−3981.94	−2.73
2b5c	2b5con1hip	0.59	−2.69	−0.01	−1.39	−5888.38	−5590.71	−3.50
2cdv	2cdvon2ssi	5.46	−8.90	**0.05**	**0.53**	−7316.75	−7011.46	−2.85
2ci2	2ci2on1sn3	5.16	−7.69	0.00	−1.26	−4340.13	−4093.71	−3.79
	2ci2on2cro	7.88	−9.44	**0.01**	−0.75	−4340.13	−4190.27	−2.31
2cro	2croon1sn3	8.18	−10.91	−0.03	−1.41	−4125.04	−3853.95	−4.17
	2croon2ci2	1.46	−3.26	−0.07	−1.18	−4125.04	−3926.64	−3.05
2cyp	2cypon1rhd	3.78	−6.33	−0.02	−1.04	−19575.27	−18517.31	−3.61
2i1b	2i1bon1lh1	2.97	−3.75	−0.03	−1.13	−9935.25	−9638.66	−1.94
2paz	2pazon1bp2	1.09	−3.12	−0.04	−1.00	−7762.56	−7385.89	−3.06
2ssi	2ssion2cdv	0.97	−0.31	−0.10	−0.98	−5459.57	−5414.86	−0.42
2tmn	2tmnon2ts1	2.58	−5.31	−0.06	**0.35**	−18767.78	−17993.23	−2.44
2ts1	2ts1on2tmn	1.24	−2.44	**0.01**	−0.85	−20658.24	−20009.35	−2.05
5pad	5padon1rei	4.53	−4.86	−0.05	−1.52	−12149.73	−11747.57	−1.90
5rxn	5rxnon1fdx	5.97	−8.70	**0.01**	−1.02	−3748.68	−3546.94	−3.74

Identifying Misfolded Proteins

We applied the CHARMM/GB force field toward the discrimination of a database containing 26 misfolded and properly folded protein structure pairs originally constructed by Holm and Sander [14]. In all but one case, a simple linear sum of nonbonded energy terms was able to distinguish between properly folded and misfolded structures (Tab. VII). The observed specificity seems to be derived primarily from the Coulomb and generalized Born electrostatic terms rather than the non-polar energy components.

The generalized Born term alone is able to identify the properly folded state in 24 of the 26 structure pairs (Tab. VII). In contrast to previous studies examining electrostatic solvation terms [9, 30], the generalized Born term seems to consistently favor the misfolded conformation. The reason for this is directly related to the anti-correlated behavior of the electrostatic solvation energy and the intra-solute Coulomb energy. Weak pairing of charged and polar residues in the misfolded conformation make these components available for interaction with solvent and thus demonstrate a more favorable GB energy. This can be substantiated by breaking down the generalized Born solvation into the self and cross polarization components.

The generalized Born solvation energy (Eq. (1)) can be simply separated into two terms. The first term is one in which the solvent polarization induced by a given charge interacts with itself. This is referred to as self polarization and occurs in Eq. (1) when $i=j$. The second term deals with charges that interact with one another ($i \neq j$) through their respective induced solvent polarization. This is referred to as the cross polarization energy and is primarily responsible for the anticorrelated behavior of the GB and Coulomb energy terms. Like charges within generalized Born intert favorably through an oppositely charged solvent polarization while unlike charges repel. This is represented in Eq. (1) by the preceding negative coefficient. Although it is clear that both the self and cross polarization terms contribute toward the identification of misfolded structures, cross polarization plays a dominant role (Tab. VIII). Simply, this means that the lack of optimal charge pairing in misfolded structures, which

TABLE VIII Percent of structure pairs that favor the folded or misfolded conformations with regard to the generalized Born self polarization energy, cross polarization energy and total GB energy. The cross polarization component, anticorrelated with the Coulomb interaction energy, strongly favors the misfolded conformations

	ΔG^{self}	ΔG^{cross}	ΔG^{total}
Folded	44%	28%	8%
Misfolded	56%	72%	92%

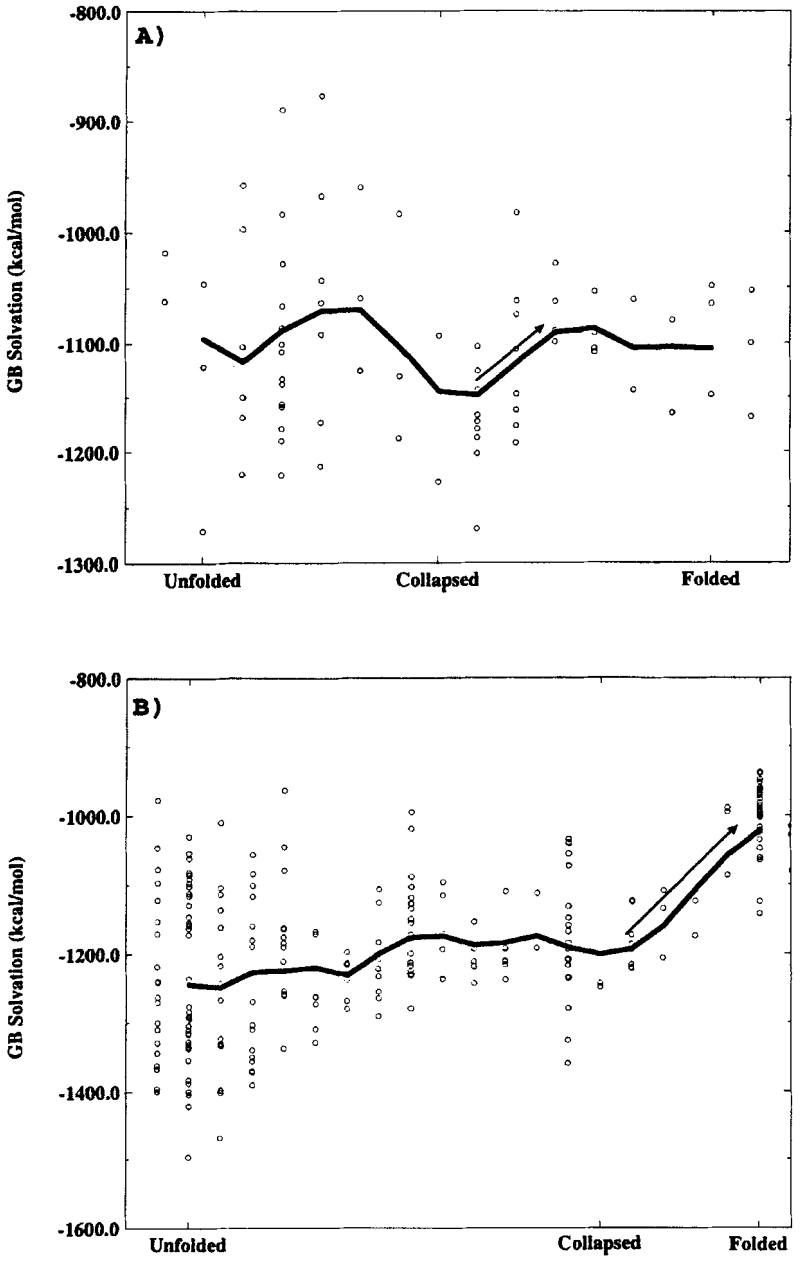

FIGURE 2 Electrostatic solvation energy profiles at various points along a folding landscape of (A) protein G and (B) CspA. Lines indicate average values at each step along the progress variable Q. The arrows indicate the solvation penalty incurred by moving from collapsed states to the final folded state.

favors the folded state in the Coulomb energy, will favor the misfolded state with regard to electrostatic solvation.

As a further test of this principle, GB solvation energies were compared in structures obtained from protein unfolding simulations of cold shock protein A (CspA) (Brooks, unpublished results) and protein G [22, 24] performed in explicit solvent. These structures represent root mean squared (rms) cluster centers from unfolding simulations grouped based on two progress variables: the radius of gyration and the number of native contacts (Q). These cluster center conformations in concert with room temperature molecular dynamics were used to construct potentials of mean force for the folding of these proteins. In this study the cluster centers were used to compute average solvation energies as a function of the progress variables. Collapsed states, identified as conformations having approximately the same radius of gyration as the folded state but with fewer native contacts, are compared to the misfolded protein structures examined from the EMBL database. Just as in the misfolded structures, collapsed conformations also tend to be more favorable in terms of solvation energy than their properly folded counterparts (Fig. 2). Since the hydrophobic solvation term is believed to drive the collapse of these proteins, this analysis raises the possibility that the solvation potential (electrostatic and nonpolar) forms a minimum at collapsed misfolded states during protein folding.

CONCLUSIONS

In this work we have demonstrated that generalized Born is able to reproduce solvation energies obtained from the finite difference solution of the Poisson equation. Further, this model has the potential to compensate for the effects of bulk solvent in the context of molecular dynamics simulations. In addition, the generalized Born solvent is highly efficient both in terms of reduced computational cost as well as its ability to utilize cutoffs without significant errors in the energy terms or force vectors.

We have also demonstrated that generalized Born may be used to rapidly identify misfolded from properly folded protein conformations. Misfolded conformations tend to have a more favorable electrostatic solvation energy due to mispairing of charged and polar amino acids. This is shown to be applicable also in the case of collapsed conformations obtained from explicit solvent unfolding simulations. The parameterization and application of the generalized Born model described in this study demonstrates that implicit solvent models not only increase the efficiency of current simulation

and screening studies, but also introduce new and insightful approaches for analysis.

References

[1] Honig, B., Sharp, K. and Yang, A. (1993). "Macroscopic models of aqueous solutions: biological and chemical applications", *J. Phys. Chem.*, **97**, 1101–1109.

[2] Roux, B. and Simonson, T. (1999). "Implicit Solvent Models", *Biophys. Chem.*, **78**, 1–20.

[3] MacKerell, A. D. J., Nilsson, L., Rigler, R. and Saenger, W. (1988). "Molecular dynamics simulations of ribonuclease T1: analysis of the effect of solvent on the structure, fluctuations, and active site of the free enzyme", *Biochemistry*, **27**, 4547–4556.

[4] Guenot, J. and Kollman, P. A. (1992). "Molecular dynamics studies of a DNA-binding protein: 2. An evaluation of implicit and explicit solvent models for the molecular dynamics simulation of the Escherichia coli trp repressor", *Prot. Sci.*, **1**, 1185–205.

[5] Norm, M., Haeffner, F., Hult, K. and Edholm, O. (1994). "Molecular dynamics simulations of an enzyme surrounded by vacuum, water, or a hydrophobic solvent", *Biophys. J.*, **67**, 548–559.

[6] Still, W. C., Tempczyk, A., Hawley, R. C. and Hendrickson, T. (1990). "Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics", *J. Am. Chem. Soc.*, **112**, 6127–6129.

[7] Born, M. (1920). "Volumes and Heats of Hydration of Ions", *Z. Phys.*, **1**, 45–48.

[8] Warwicker, J. and Watson, H. C. (1982). "Calculation of the Electric Potential in the Active Site Cleft due to α-Helix Dipoles", *J. Mol. Bio.*, **157**, 671–679.

[9] Gilson, M. K. and Honig, B. (1988). "Calculation of the Total Electrostatic Energy of a Macromolecular System: Solvation Energies, Binding Energies, and Conformational Analysis", *Proteins*, **4**, 7–18.

[10] Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. and Karplus, M. (1983). "CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations", *J. Comp. Chem.*, **4**, 187–217.

[11] Skolnick, J., Kolinski, A. and Ortiz, A. R. (1998). "Reduced protein models and their application to the protein folding problem", *J. Biomol. Struct. Dyn.*, **16**, 381–396.

[12] Shakhnovick, E. I. (1997). "Theoretical studies of protein folding thermodynamics and kinetics", *Curr. Opin. Struct. Biol.*, **7**, 29–40.

[13] Karplus, M. and Sali, A. (1995). "Theories of protein folding", *Curr. Opin. Struct. Biol.*, **5**, 58–73.

[14] Holm, L. and Sander, C. (1992). "Evaluation of Protein Models by Atomic Solvation Preference", *J. Mol. Bio.*, **225**, 93–105.

[15] Jackson, J. D., "*Classical Electrodynamics*", John Wiley and Sons, Inc., New York, 1999.

[16] Qiu, D., Shenkin, P. S., Hollinger, F. P. and Still, W. C. (1997). "The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii", *J. Phys. Chem. A*, **101**, 3005–3014.

[17] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P., "*Numerical Recipes in C*", Press Syndicate of the University of Cambridge, New York, 1992.

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

[19] MacKerell, A. D. Jr., Wiorkiewicz-Kuczera, J. and Karplus, M. (1995). "An All-Atom Empirical Energy Function for the Simulation of Nucleic Acids", *J. Am. Chem. Soc.*, **117**, 11946–11975.

[20] 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 III, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D. and Karplus, M. (1998). "All-atom

Empirical Energy Function for Molecular Modeling and Dynamics Studies of Proteins", *J. Phys. Chem. B*, **102**, 3586–3616.

[21] Ryckaert, J. P., Ciccotti, G. and Berendsen, H. J. C. (1977). "Numerical integration of the Cartesian equations of motion for a system with constraints: molecular dynamics of *n*-alkanes", *J. Comp. Phys.*, **52**, 251–292.

[22] Sheinerman, F. B. and Brooks III, C. L. (1998). "Calculations on Folding of Segment B1 of Streptococcal Protein G", *J. Mol. Bio.*, **278**, 439–456.

[23] Sheinerman, F. B. and Brooks III, C. L. (1997). "A molecular dynamics simulation study of segment B1 of protein G", *Proteins*, **29**, 193–202.

[24] Sheinerman, F. B. and Brooks III, C. L. (1998). "Molecular Picture of Folding of a Small Alpha/Beta Protein", *PNAS*, **95**, 1562–1567.

[25] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). "The Protein Databank: A computer based archival file for macromolecular structures", *J. Mol. Bio.*, **112**, 535–542.

[26] van Gunsteren, W. F. and Berendsen, H. J. C. (1987). "*GROMOS, Groningen Molecular Simulation Computer Program Package*".

[27] Lazaridis, T. and Karplus, M. (1999). "Discrimination of the Native from Misfolded Protein Models with an Energy Function Including Implicit Solvation", *J. Mol. Bio.*, **288**, 477–487.

[28] Sitkoff, D., Sharp, K. A. and Honig, B. (1994). "Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models", *J. Am. Chem. Soc.*, **98**, 1978–1988.

[29] Edinger, S. R., Cortis, C., Shenkin, P. S. and Friesner, R. A. (1997). "Solvation Free Energies of Peptides: Comparison of Approximate Continuum Solvation Models with Accurate Solution of the Poisson-Boltzmann Equation", *J. Phys. Chem. B*, **101**, 1190–1197.

[30] Vorobjev, Y. N., Almagro, J. C. and Hermans, J. (1998). "Discrimination Between Native and Intentionally Misfolded Conformations of Proteins: ES/IS, a New Method for Calculating Conformational Free Energy That Uses Both Dynamics Simulations With an Explicit Solvent and an Implicit Solvent Continuum Model", *Proteins*, **32**, 399–413.