New Analytic Approximation to the Standard Molecular Volume Definition and Its Application to Generalized Born Calculations

MICHAEL S. LEE, MICHAEL FEIG, FREDDIE R. SALSBURY, JR.,* CHARLES L. BROOKS III

Department of Molecular Biology (TPC 6), The Scripps Research Institute,
10550 North Torrey Pines Road, La Jolla, California 92037

Received 1 October 2002; Accepted 8 January 2003

Abstract: In a recent article (Lee, M. S.; Salsbury, F. R. Jr.; Brooks, C. L., III. J Chem Phys 2002, 116, 10606), we demonstrated that generalized Born (GB) theory provides a good approximation to Poisson electrostatic solvation energy calculations if one uses the same definitions of molecular volume for each. In this work, we present a new and improved analytic method for reproducing the Lee–Richards molecular volume, which is the most common volume definition for Poisson calculations. Overall, 1% errors are achieved for absolute solvation energies of a large set of proteins and relative solvation energies of protein conformations. We also introduce an accurate SASA approximation that uses the same machinery employed by our GB method and requires a small addition of computational cost. The combined methodology is shown to yield an efficient and accurate implicit solvent representation for simulations of biopolymers.

© 2003 Wiley Periodicals, Inc. J Comput Chem 24: 1348–1356, 2003

Key words: molecular surface area; electrostatic calculations; Poisson–Boltzmann; hydrophobic effect; molecular mechanics

Introduction

At present, the majority of biomolecular simulations are performed with an explicit representation of the solvent molecules. Recent progress, however, has been made with implicit solvent models, such as generalized Born (GB).[1–8] The GB model is an approximation to the Poisson electrostatic solvation model, which is a continuum dielectric model of the solvent. First, GB calculations should be able to match Poisson results. However, the Poisson solvation energy of a set of point charges is a unique quantity only insofar as the boundary conditions are defined, that is, the dielectric boundary of the solute, and this is related to the definition used for the solute volume.

The most naïve definition of the solute volume is a superposition of atomic van der Waals (vdW) spheres, also known as the vdW volume.[9] This definition produces small high-dielectric cavities in the interior of a solute. This is an artifact that necessarily leads to overestimation of the solvation energy. Alternatively, in the solvent-accessible volume model, the vdW spheres are augmented by a probe radius, for example, 1.4 Å, to remove these problematic high-dielectric gaps. However, the resultant surface of the solute becomes overextended by the probe radius, which leads to the solvation energy being instead systematically underestimated. A compromise between these two approaches, known as the molecular surface,[10] is to start with the solvent-accessible model and remove the volume along the surface by carving out water probes whose centers lie on that surface. The resulting volume, which is enclosed by the molecular surface, shall be termed the standard molecular volume (SMV). It has a boundary similar to that of the vdW volume without the problematic interior high-dielectric gaps. The SMV definition would be the solute volume of choice if it were not for one significant drawback, that is, that it does not have smooth differentiability with respect to atomic displacements for certain configurations of atoms.[11] This problem would lead to unphysically large forces in a molecular dynamics simulation.

Nonetheless, almost all GB methods in the literature attempt to mimic SMV-based Poisson solvation energies. The GB energy is a sum of two contributions: the self-polarization of each individual atom and the cross-polarization of all pairs of atoms. The cross-polarization of two atoms is uniquely defined, in GB models, by three values: the distance between the two atoms and the self-polarization energies of each atom. Thus, the key to getting good

Correspondence to: C. L. Brooks III; e-mail: brooks@scripps.edu

*Present address: Department of Physics, Wake Forest University, Winston-Salem, NC 27106.

agreement between GB and Poisson results lies in the accurate calculation of the atomic self-polarization energies. It has been shown in a previous article[8] that given accurate self-polarization energies the GB model can reproduce Poisson solvation energies to ≈1% accuracy. Several GB approaches that are designed to obtain atomic self-polarization energies have been presented in the literature. They can be roughly divided into two classes: atomic pair summations[2–5] and surface/volume integrations.[1,6–8] All of the atomic pair-based methods are analytic with respect to atomic position and can be used in energy minimizations and molecular dynamics simulations. In most of the GB surface/volume integration literature, on the other hand, gradients with respect to atomic positions have yet to be implemented. The one exception, so far, is the analytic GB method of Lee et al., which performs numerical integration over an analytically defined volume.

Arguably, the most accurate GB algorithms are the volume grid-based method of Lee et al. and the surface grid-based approach of Ghosh et al.[7] Both of these schemes utilize the SMV definition to mimic SMV-based Poisson calculations. In Lee et al., for example, atomic self-polarization energies are obtained by volume integration over the SMV and the addition of a simple empirical correction to the Coulomb field approximation. Both absolute solvation energies of a large protein dataset and relative solvation energies for two sets of protein conformations were obtained with ≈1% accuracy compared to the benchmark Poisson results. Further, the method was rigorous in that it required only two adjustable parameters. Nevertheless, neither the method of Lee et al. nor the method of Ghosh et al. have well-defined gradients with respect to atomic position. This is because of the above-stated problems with the underlying SMV definition.

One way to circumvent this issue might be to redefine the solute volumes for benchmark Poisson calculations. An alternative volume definition, which allows for differentiability, is a superposition of atomic spherically symmetrical functions.[12,13] In these models, however, a compromise must be made between achieving a good surface definition and minimizing interior voids. Thus, while such models provide robust definitions in their own right, they are not as physically justifiable as the original SMV definition.

The primary goal of this work is to introduce a new analytic volume model for GB that closely approximates the SMV definition yet also allows for smooth differentiability with respect to atomic displacements. This model is based on a superposition of atomic functions with one key twist: a vector-based scaling term that corrects the usual discrepancies that occur in a superposition approach. The general problem of matching superposition volumes to SMV can be seen for the case of two atoms in Figure 1. Usually, the size of the atomic functions is a balance between trying to get the correct volume for two touching atoms and for two nontouching, but nearby, atoms. As can be seen in Figure 1, if large radial functions are used to fill in the void between nearby atoms the model will produce an unphysical bulge when the atoms are touching or interpenetrating. On the other hand, if small radial functions are used to get the right volume for two touching atoms then the model will have a void when the atoms are slightly apart. The solution to this problem is to consider the vectors from a volume point of interest to the two atomic centers. The sum of the two vectors will add constructively if the volume point is off-axis. On the other hand, the two vectors will add destructively if the volume point is near the axis. Thus, vector summation is a simple criterion to differentiate between off-axis and near-axis regions. It follows that our novel scaling term is an inverse measure of this vector summation, such that the near-axis regions can be amplified and the off-axis regions can be attenuated. This technique also works well for clusters of more than two atoms. With this new analytic approximation to the SMV definition, an *analytic* GB method is presented here that can reproduce SMV-based Poisson solvation energies to nearly the same degree of accuracy as the original grid-based approach.

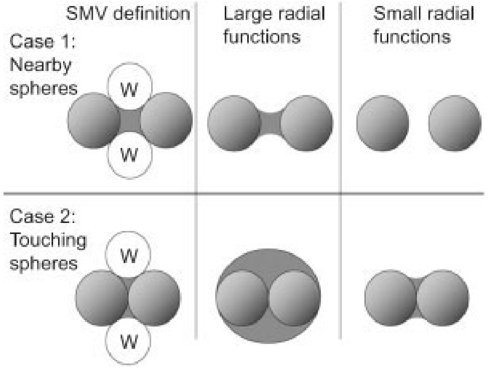

Figure 1. Schematic diagram comparing the SMV definition to volumes obtained by superposition of atomic functions with large and small radial extents. The cases illustrated here consist of two equally sized atoms. The circles with the letter "W" inside of them indicate water probes.

It should be pointed out that analytic gradient calculations have been attempted with both surface-mesh and volume-grid Poisson methods.[11,14,15] These Poisson methods are appropriate for energy minimizations on small molecules. However, for larger systems it is not clear, given the exact SMV definition, that the derivatives will be finite for all possible configurations of atoms. For example, as a gap is formed in the interior of a solute there will come a point when a water probe will suddenly be able to fit. At this point, the forces will become infinite.[11]

Using the same machinery employed in our analytic methods, we can also calculate an approximation to the solvent-accessible surface area (SASA) with little additional computational cost. Our new surface area methodology is at least as accurate as most fast pair-based schemes.[16,17]

In the remainder of this article, the original analytic GB scheme of Lee et al. and the new GB method are discussed. In the Results section, the differences between the solute volumes used in the old and new methods are illustrated graphically. Also, accuracies of solvation energies for the old and new methods are compared to the grid-based approach. Further, our new approximate surface area scheme is evaluated against exact analytic results. Finally, we summarize and discuss future directions.

Theory and Implementation

The GB molecular volume (GBMV) method was originally detailed in Lee et al.[8] The methodology has been improved in several respects. Here, we give an overview of the original analytic model and detail the differences between our old and new approach.

In modern GB theory, devised by Still and coworkers,[1] the implicit electrostatic solvation energy of a solute with dielectric, $\varepsilon_{\text{solute}}$, and solvent, $\varepsilon_{\text{solvent}}$, is a summation over atom pairs,

$$E_{\text{GB}} = k\sum_{i,j}\frac{q_iq_j}{\sqrt{r_{ij}^2 + \alpha_i\alpha_j\exp(-r_{ij}^2/K_s\alpha_i\alpha_j)}} \tag{1}$$

where indices i and j loop over atoms, q_i is the atomic charge on atom i, r_{ij} is the distance between atoms i and j, and $k = -166.0^*$ $(\varepsilon_{\text{solute}}^{-1} - \varepsilon_{\text{solvent}}^{-1})$ when the energy is expressed in kcal/mol and the units of distance are Å. The Born radii of the atoms, α_i, are inversely proportional to the atomic self-polarization energies, $\alpha_i = kq_i^2/G_i^{\text{pol}}$ [see eq. (2)]. The constant, K_s, in the original Still equation is equal to 4. Other possible values have been suggested in the literature.[2,18] We found $K_s = 8$ to be nearly optimal for calculation of solvation energies.[19] Regardless of the K_s value used, given an accurate calculation of the Born radii, the GB solvation energy can achieve a mean unsigned error of 1% with respect to the Poisson energy.[8] Thus, the crux of the problem in GB theory is the calculation of the Born radii and hence the atomic self-polarization energies. It was shown in Lee et al. that given a faithful representation of the molecular volume, V, one could derive accurate self-polarization energies, G_i^{pol}, using the following empirical formula:

$$G_i^{\text{pol}} = -k\left(\frac{1}{R} - \frac{1}{4\pi}\iiint\limits_{|\vec{r}-\vec{x}_i|>R}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^4}\,dxdydz\right) + kP\sqrt{\frac{1}{2R^2} - \frac{1}{4\pi}\iiint\limits_{|\vec{r}-\vec{x}_i|>R}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^5}\,dxdydz} \tag{2}$$

where R is a value less than or equal to the vdW radius of atom i, R_i. The first term in eq. (2) is the *negative* of the Coulomb field approximation and the second term is an empirically derived correction term. The constant, P, should equal $2\sqrt{2}$ to achieve the correct single atom limit; however, slightly different values were used in Lee et al. as a results of parameter optimization. Another alternative formula, which we will henceforth use in our new GB model, has also been empirically obtained and achieves a modestly better fit to Poisson radii:

$$G_i = k\left(1 - \frac{1}{\sqrt{2}}\right)\left(R_i^{-1} - \frac{1}{4\pi}\int\int_{R_i}^{\infty}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^4}\,drd\Omega\right) + k\left(\frac{1}{4R^4} - \frac{1}{4\pi}\int\int_{R_i}^{\infty}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^7}\,drd\Omega\right)^{1/4} \tag{3}$$

It is likely that a general equation of the form

$$G_i^{\text{pol}} = k\left(1 - \frac{1}{\sqrt{2}}\right)\left(\frac{1}{R} - \frac{1}{4\pi}\iiint\limits_{|\vec{r}-\vec{x}_i|>R}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^4}\,dxdydz\right) + k\left(\frac{1}{4R^4} - \frac{1}{4\pi}\iiint\limits_{|\vec{r}-\vec{x}_i|>R}\frac{V(\vec{r})}{|\vec{r}-\vec{x}_i|^7}\,dxdydz\right)^{1/4} \tag{4}$$

may be used to obtain even closer fits to Poisson radii. Each term has the correct dimensions of energy. However, the actual coefficients of this summation, at present, must be derived empirically through fitting. Also, it is not clear whether an exact solution to the problem can be achieved with this particular form.

Solute Volume Definitions

The solute volume in Poisson solvation energy calculations that we intend to mimic is the standard molecular volume (SMV) with a water probe radius of 1.4 Å. This volume is usually constructed in two steps: First, a solvent-accessible volume is constructed by the superposition of atomic spheres with radii that are 1.4 Å larger than their van der Waals radii. Then, 1.4-Å water probes with centers along the solvent-accessible surface are carved out of this solvent-accessible volume.

In Lee et al., two GB models were introduced. The first model entailed building the SMV volume on an arbitrarily precise numerical grid. This led to the accurate calculation of atomic self-polarization energies. The other model, which will be denoted GB-MV1, used an analytic approximation to the SMV and was less successful at accurately reproducing self-polarization energies. For GB-MV1, a preprocessed molecular volume is built up as a superposition of atom-centered quartic exponentials*:

$$A_j(\vec{t}_j) = \exp(\gamma_j\|\vec{t}_j\|^4),\ S_{\text{MV1}}(\vec{r}) = \sum_j A_j(\vec{t}_j), \quad \text{with } \gamma_j = \frac{\gamma_0\ln(\lambda)}{R_j^4} \tag{5}$$

where R_j is the vdW radius of atom j, $\vec{t}_j = \vec{r} - \vec{x}_j$, $\vec{x}_j$ is the center of the atom j, and γ_0 and λ are adjustable parameters. This preprocessed molecular volume, $S_{\text{MV1}}(\vec{r})$, is then quenched, via a Fermi–Dirac switching function,[20] to produce a function V, which is bounded between 0 and 1,

$$V_{\text{MV1}}(\vec{r}) = \frac{1}{1 + \exp(\beta(S_{\text{MV1}}(\vec{r}) - \lambda))} \tag{6}$$

where the adjustable parameter, β, controls the width of the switching function and λ determines the midpoint. The GB-MV1 method performs satisfactorily for absolute solvation energies compared to Poisson calculations. However, another conventional analytic GB model[5] produces slightly better relative solvation energies across a range of protein conformations when an implicit

*In Lee et al., we had a typographical error in the equation that is correct here.

hydrogen force field, PARAM19,[21] is used. The primary reason that this model does not work well was that a unique parameterization to find balance between touching atoms and nontouching, but nearby, atoms could not be found.

To clarify the derivation of a closer approximation to the SMV, two definitions are needed. A "gap region" is an open space *inside* a cluster of atoms. On the other hand, an "open" region is an open space *outside* a cluster of atoms. The main problem in superposition methods is that even when an optimal-sized atomic function is obtained the gap regions tend to be underfilled and the open regions tend to bulge. Thus, what is needed is a means to identify these two regions and correct their respective behaviors.

Consider the vectors joining, $\vec{t}_j$, the point of interest, $\vec{r}$, and each atomic center, $\vec{x}_j$. Vectors that point to a gap region will typically add destructively, while vectors that point to an open region will usually add constructively. Thus, the resultant vector is a means to identify gap and open regions. Accordingly, our new model, which we will call MV2, enhances the gap regions and diminishes the open regions through the use of a scaling factor that is an *inverse* measure of the normalized resultant vector just described:

$$S_{\mathrm{MV2}}(\vec{r}) = S_0\left[\sum_j F_{\mathrm{MV2}}(\|\vec{t}_j\|)\right]\frac{\sum_j \|\vec{t}_j\|^2 F^2_{\mathrm{MV2}}(\|\vec{t}_j\|)}{\left\|\sum_j \vec{t}_j F_{\mathrm{MV2}}(\|\vec{t}_j\|)\right\|^2}, \qquad (7)$$

where S_0 is an adjustable empirical parameter and F_{MV2} is a spherically symmetrical atomic function described below. The first bracketed term is a superposition of atomic volumes similar to the original definition, S_{MV1}. The second term, however, is a scaling factor in which the numerator is a sum of the magnitudes of each vector and the denominator is the magnitude of the resultant sum of the vectors. The numerator acts as a normalization term for the denominator. Note that the vectors are weighted by the atomic functions. This assures that only the relevant vectors contribute to the determination of the scaling factor. Further, it should be noted that large values of S_{MV2} that may result from small denominators are quenched by the switching function defined above. Denominators below a certain threshold, for example, 10^{-14}, are pegged to 10^{-14} for numerical purposes only.

Given that we are trying to reproduce the SMV as closely as possible, we re-evaluated the use of various atomic functions. There are a few guidelines for choosing an optimal atomic function for this problem. First, the function should have a value greater than or equal to one inside the vdW sphere. Second, the tail of the function should decay monotonically from one at the vdW surface to approximately zero at about 2.8 Å from the spherical surface. Third, the tail should have a similar decay length regardless of the size of the atom. One of the most obvious functions that satisfies these three criteria is the exponential function, $F(\vec{t}_j) = \exp(-\alpha[\|\vec{t}_j\| - R_j])$. It turns out that satisfactory results can be obtained with a single value of $\alpha = -1.98$ for all atomic radii. However, because we desired to make the atomic function as computationally inexpensive as possible, we decided to avoid the exponential and square root evaluations that would be necessary with this function. Instead, we chose the function

$$F_{\mathrm{VSA}}(\vec{t}_j) = \frac{C_j^2}{[C_j + \|\vec{t}_j\|^2 - R_j^2]^2},\ C_j = P_1R_j + P_2 \qquad (8)$$

where P_1 and P_2 are empirically fitted parameters. This function does a satisfactory job in mimicking the tail of the exponential. Because there are now two parameters, this function can actually be fit to provide slightly better results than the exponential. In practice, we smoothly truncate the tail of this function to zero in the range of 1.9–2.1 Å using the polynomial switching function, *f*, which is described below. This truncation improves computational speed somewhat without significantly affecting accuracy. Actually, truncation can improve the quality of the MV2 model at large internuclear separations where spurious features may arise (see Results).

The previous analytic function, $S_{\mathrm{MV1},}$ monotonically increases as atoms are summed into the formula. This means that when the summation reaches a threshold value the summation can be immediately terminated. In the new model, however, the summation over atoms cannot be terminated early because the vector-based scaling term is nonmonotonic. For this reason, the new model has approximately double the computational cost of the MV1 method. One way to reduce computational cost is to probe a vdW volume before evaluating the MV2 function. At the least, the vdW volume is a subset of the SMV. Thus if a point of interest, $\vec{r}$, is inside the vdW volume, the more complicated MV2 function need not be evaluated. The vdW volume, in our scheme, is a superposition of switching functions, *f*, with vdW radial extents and short tails. The mathematical definition of our vdW volume is as follows:

$$S_{\mathrm{vdW}}(\vec{r}) = 2\sum_j f(u_j) \qquad (9)$$

where

$$u_j = \frac{\|t_j\|^2 - (R_j + t_-)^2}{(R_j + t_+)^2 - (R_j + t_-)^2} \qquad (10)$$

and

$$f(u) = \begin{bmatrix} 1 & u \le 0 \\ 1 - 10u^3 + 15u^4 - 6u^5 & 1 > u > 0 \\ 0 & u \ge 1 \end{bmatrix} \qquad (11)$$

where t_- and t_+ denote the minus and plus widths of the atomic vdW function, respectively. The factor of 2 in eq. (9) guarantees the vdW function will drop off starting at the vdW surface. After some testing, the following values have been assigned: $t_- = -0.125$ Å and $t_+ = 0.25$ Å. The fifth-order polynomial, *f*, was chosen because it has continuity up to second derivatives at the $u = 0$ and $u = 1$ boundaries. The vdW volume, $S_{\mathrm{vdW},}$ is added to the preprocessed MV2 volume, S_{MV2}, before the switching function [eq. (6)] is applied. To guarantee smoothness in the molecular volume in the region where the MV2 model and the vdW volume intersect, the MV2 atomic function, F_{MV2}, is switched on using the scaling factor, $1 - f$, in reverse of the vdW volume being turned off. As mentioned above, the truncation of the MV2 atomic function is achieved by scaling with the same function, $f(u_j)$, except that for this case $t_+ = 1.9$ Å and $t_- = 2.1$ Å.

In our previous model, MV1, we transformed the radii with a shifting factor, α_0. We also modified the coefficient, P, in the second term of eq. (2) from its ideal value. In our new model, we use eq. (3) instead and do not alter its coefficients. Instead, we apply a linear transformation, $\alpha_1 = C_1\alpha_i + C_0$, to improve the fit of the MV2 radii to Poisson-derived Born radii.

Supporting Algorithms

For any choice of volume function, the integrals in eqs. (2) and (3) are approximated by standard numerical quadrature techniques that were detailed in Lee et al.[8] To briefly summarize, however, the integrals are split up into radial and angular components. The angular integral was approximated in the MV1 scheme as a summation over the 20 vertices of a dodecahedron. However, in the new scheme the 38-point Lebedev angular grid[22] is used because it provides improved accuracy. Further, alternating between 0°/90° Z-axis rotation of the angular grid at successive radial points improves results with no added cost. The radial integral is approximated by a Riemann–Stieltjes summation[23] of the form

$$\int_{R_o}^{R_{\mathrm{MAX}}} r^2 \frac{V(r)}{r^n}\, dr = \frac{1}{n-3} \sum_{w=1}^{m-1} V\!\left(\frac{q_w + q_{w+1}}{2}\right)\left[\frac{1}{q_w^{n-3}} - \frac{1}{q_{w+1}^{n-3}}\right] \tag{12}$$

The integration points, q_m, have been modified from Lee et al. to achieve better efficiency and accuracy. The following new series (in Å) is used: {0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0, 6.0, 8.0, 10.0, 14.0, 18.0}. The number of integration points necessary for a particular atom varies with the vdW radius of the atom. The first radial point needed for an atom is the largest value smaller than the vdW radius of that atom.

For both analytic models, GB–MV1 and GB–MV2, a lookup table needs to be generated so that at each integration point the atoms contributing to the solute volume function at that point can be quickly determined. The algorithm for generating a lookup table was first described in Lee et al. We repeat the description here because we have updated the approach and made a few changes specifically for the MV2 model. The lookup table spans a spatially uniform cubic grid. At each grid point, all the atoms that can possibly contribute to the cubic cell are stored. An atom contributes to a grid cell if it is less than a distance, $r_{\max}$, from the center of the cell, where $r_{\max}$ is defined as

$$r_{\max} = R_i + 2.1 + \frac{\sqrt{3}}{2}c + r_{\mathrm{buffer}} \tag{13}$$

where the parameter, c, is the width of the grid cells, the value 2.1 Å is the length of the tail of the MV2 atomic function as prescribed above, and r_{buffer} is a user-adjustable length that determines how far any atom can move from its initial position before the entire lookup table needs to be rebuilt.

There are two steps in our lookup table method. The first procedure is a loop over all atoms, where each atom is "painted" onto the grid points to which it could contribute. Two passes over all atoms are made. The first pass counts the number of atoms that contribute to each grid cell so that the right amount of memory can be allocated to each grid cell. The second pass actually generates the list of atoms per grid cell. After the painting process is complete, the second step is to sort the relevant atoms for each grid cell based on how relevant they are to that cell. The criterion that we use for an atom i is $1 - d^2/R_i^2$, where d is the distance between the cell center and atom center. Larger values of the criterion indicate that an atom is more relevant to a cell. The sorting procedure reduces the amount of work performed by the numerical integration procedure because if an integration point taps the interior of the vdW volume function the summation for that point can be immediately terminated.

There are certain computational issues to consider with the lookup table procedure. First, the lookup table method requires a certain amount of computational time. Thus, it is probably not desirable to build a lookup table every time the Born radii are computed. Increasing the value of r_{buffer} reduces the frequency of rebuilding the lookup table. However, it also leads to more memory usage and slower integration because the number of atomic neighbors per integration point is increased. We have found that values of r_{buffer} from 0.5–1.0 Å provide a reasonable balance of time and memory.

SASA Approximation

With the numerical integration and lookup table algorithms in place, it is relatively easy to implement an approximate analytic calculation of the solvent accessible surface area.[9,24] Consider each atom, i, to have an enlarged spherical function with a radius equal to $R_i + 1.4$ Å. Then, in a method we will denote SASA-1, we find the approximate exposed surface area of atom i by scanning the enlarged surface through a series of angular integration points. If an integration point is inside the spherical function of another atom, then it is considered buried. The exposed surface area of atom i, σ_i, is the total surface area of the enlarged sphere minus the weighted angular points that are buried:

$$\sigma_i = 4\pi(R_i + 1.4)^2 - \sum_m w_m f\left[\sum_j f(u_{mij})\right], \quad \text{with } \sigma = \sum_i \sigma_i \tag{14}$$

where

$$u_{mij} = \frac{\|\vec{r}_m + \vec{x}_i - \vec{x}_j\|^2 - (R_j + t_-^{SA})^2}{(R_j + t_+^{SA})^2 - (R_j + t_-^{SA})^2}, \tag{15}$$

σ is the total surface area of the system, w_m are angular integration weights, the widths, t_-^{SA} and t_+^{SA}, have been chosen to be 1.2 and 1.5 Å respectively, and f has the same definition as eq. (11). Other values of t_-^{SA} and t_+^{SA} can also be used. However, the values that we chose were the best compromise between obtaining good atomic surface areas and satisfactory total surface areas. If there were an infinite number of angular points on each atom and the values of

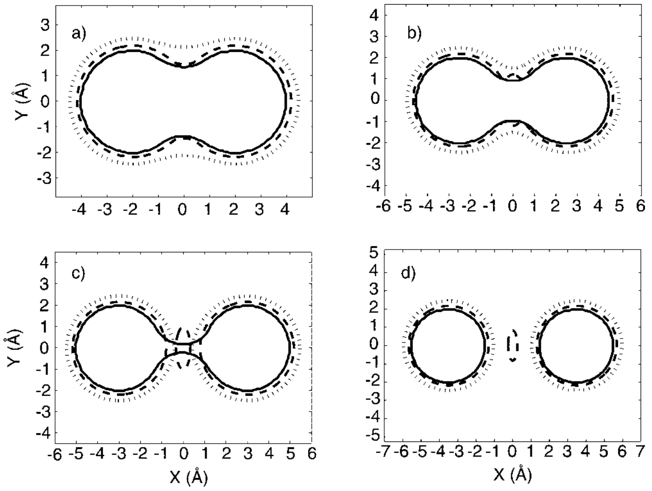

Figure 2. Contour diagrams of a planar slice through various volume definitions for 2-Å spheres (e.g., aliphatic hydrogens in the CHARMM PARAM22 force field[25]) separated by various internuclear distances, *d*: (a) *d* = 4 Å (touching), (b) *d* = 5 Å, (c) *d* = 6 Å, and (d) *d* = 7 Å. Solid lines, SMV; broken lines, MV2; dotted lines, MV1.

t_-^{SA} and t_+^{SA} were set equal to zero, the exact SASA result would be obtained. In practice, however, the 38 angular points that are also used in the GB calculation provides a modest error versus an exact SASA calculation and provides an algorithm that is many times faster than the exact analytic SASA methods currently implemented in CHARMM.[21]

Results and Discussion

The benchmarks for our GB models are Poisson electrostatic solvation energy calculations using the SMV definition with a probe radius of 1.4 Å. All Poisson calculations were performed in CHARMM[21] using the PBEQ module with a grid spacing of 0.25 Å. All of the vdW radii and atomic charges for the protein structures considered here are from the PARAM22[25] force field.

The MV1 model results presented here use the same parameters that were used in Lee et al.: $\beta = -100$, $\lambda = 0.1$, $\gamma_0 = 0.44$, $P = 2.92$, $C_0 = -0.5$, and $C_1 = 1.0$. The GB–MV2 method was implemented in a development version of CHARMM, c29a2. All of the MV2 model parameters used for this study were derived through trial and error and are optimized for the PARAM22[25] force field. It is expected that these parameters are robust with respect to force field variation, analogous to MV1. Only minor modifications of some of the parameters would be necessary to obtain optimal results for other force fields. The benchmark test set that we used for fitting was the same small protein test set that was used in Lee et al. plus a few additional proteins and a few conformations of protein L (2PTL)[26] and the villin headpiece (1VII).[27] The parameters that have not been already explicitly documented above are as follows: $\beta = -20$, $\lambda = 0.5$, $S_0 = 0.7$, $P_1 = 0.45$, $P_2 = 1.25$, $C_0 = -0.102$, and $C_1 = 0.9085$. The value of β that we have chosen is a compromise between smoothness of the solute volume function and computational speed. Values of S_0 in the range of 0.6–0.9 appear to give reasonable results when $\lambda = 0.5$. Other values of λ can be used with appropriate modifications to S_0 because these two parameters are closely related. The grid-based approach was reparameterized with the new correction factor [eq. (3)]. The two parameters of the grid-based approach are $C_1 = 0.9026$ and $C_0 = -0.008$. These parameters were derived by a least-squares fit to the Poisson-derived Born radii of the original small protein set used in Lee et al. The grid-based approach used a grid spacing of 0.2 Å and similar integration parameters as outlined in Lee et. al, except for two modifications: First, the number of ϕ angles in the angular integration grid was increased from 8 to 10; second, a smoothing filter with a 3 × 3 × 3 kernel was applied to the grid volume in the final step.

To illustrate how our new analytic volume scheme, MV2, compares to our old model, MV1, we created several simple model systems and generated 2D contour plots of a planar slice through each system. The SMV graph was obtained by our grid-based GB approach with a grid spacing of 0.1 Å and a contour value of 0.5. A contour value of 0.1 was used for MV1 and a value of 0.5 was used for MV2. In Figure 2, two spheres with radii equal to 2 Å are arranged at four different interatomic distances, *d*. The *d* = 4 Å case might correspond to a van der Waals contact between two carbon atoms. One can see that in the *d* = 4 Å and *d* = 5 Å cases MV2 obtains a closer fit to SMV than MV1. In both of these cases, the excess interatomic volume of MV1 leads to a systematic overprediction of Born radii that has to be shifted downward by the parameter, $C_0 = -0.5$ Å. MV2, on the other hand, has the right interatomic behavior and requires much less compensation, $C_0 = -0.102$ Å. However, in the *d* = 6 Å and *d* = 7 Å cases MV1 has more or less the correct behavior, although it fails to produce a bridge in the *d* = 6 Å case. On the other hand, MV2 introduces interatomic artifacts at these two distances that would presumably induce slightly larger than desired Born radii, especially for the

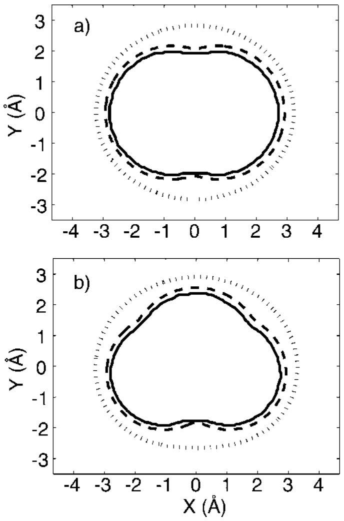

Figure 3. Contour diagrams of a planar slice through various volume definitions for typical chemical arrangements: (a) C—C single bond; (b) CO_2 carboxylate group. Solid lines, SMV; broken lines, MV2; dotted lines, MV1.

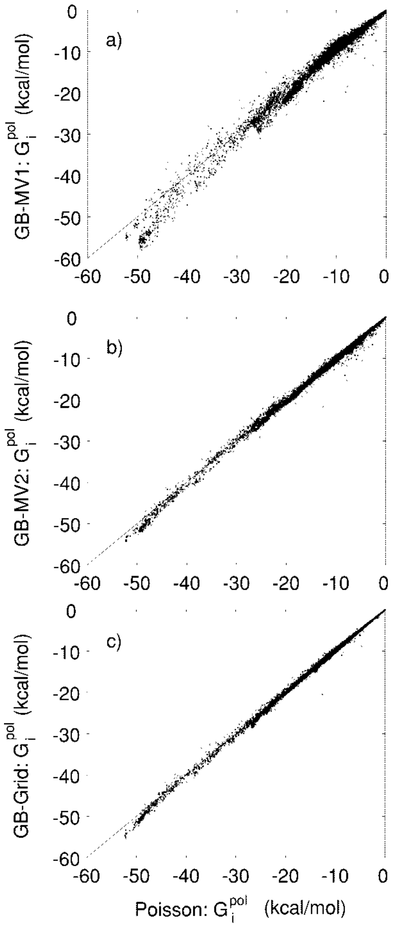

Figure 4. Atomic self-polarization energies of (a) MV1, (b) MV2, and (c) grid- versus Poisson-derived results for the 20 small proteins data set. The correlation coefficients for the MV1, MV2, and grid-based methods are 0.995, 0.9992, and 0.9996, respectively. Lines of y = x are displayed for comparison.

d = 7 Å case. In the d = 6 Å case, this artifact may provide a qualitative compensation for the bridge that is seen in the SMV contour. The artifact in MV2 is nonetheless present up to interatomic distances of ≈9 Å and is due to the long-range tail of the MV2 atomic function. Consider that in the artificial volume region the two vectors emanating from the atomic centers nearly completely cancel each other out. This leads to a small denominator in eq. (7) and hence a large scaling term. Perhaps, the artifact could be made less prominent by shortening the tail of the MV2 atomic function.

In Figure 3, two common moieties are presented: the C—C single bond and the carboxylate group, CO_2^-. In these two cases, the MV2 model performs significantly better than MV1. The MV2 contours track the SMV contours well. On the other hand, the MV1 contours have noticeable bulges that will lead to overprediction of Born radii. The atomic functions in MV1 could presumably be made smaller to reduce these bulges. However, this fix would eliminate the good results obtained for the d = 5 Å and d = 6 Å cases. The compromise between achieving good results for bonded atoms as well as nearby atoms is one of the key problems with solute volumes generated by a simple superposition of atomic functions.

Table 1. Absolute Percentage Errors in Electrostatic Solvation Energy Versus Benchmark SMV–Poisson (0.25 Å).

	MV2 (4)	MV2 (8)	Grid (8)	MV1 (4)
Absolute avg.	1.66%	1.15%	0.98%	3.06%
SD	1.35%	0.91%	0.83%	1.95%
Max	10.38%	5.22%	5.52%	12.11%

PARAM22 vdW radii and atomic charges are used. The value of K_s is denoted in parentheses. The two values of K_s are presented for MV2 so that one can differentiate between improvements in the self-polarization energies and the Still equation. Test on 611-protein structure, a diverse set (M. Feig et al., a work in progress).

With a more accurate volume model than MV1, we are now able to achieve a closer match to Poisson-derived self-polarization energies, approaching the accuracy of the grid-based procedure. Figure 4 illustrates how the MV2 technique is superior to the MV1 method for obtaining self-energies. In this illustration, it appears that the MV2 method is almost as accurate as the grid-based approach. This accuracy is consistent with the results of the contour plots described above.

Given accurate self-polarization energies, the GB-MV2 model performs well for absolute and relative solvation energies. In Table 1 and Figure 5, one can see that the MV2 method with a Still factor, K_s = 8, incurs errors about twice as small as MV1. The MV1 method was intended to perform equally well with PARAM19 and PARAM22; thus, it could probably be optimized a bit better for PARAM22. As seen in Tables 2 and 3, relative solvation energies for the MV2 model are also about twice as good as the MV1 method. Again, the MV2 model is about as good as the grid-based method. As a side note, results that are not shown here indicate the MV2 model has a similar accuracy for absolute and relative solvation energies as the Poisson method with 0.5-Å grid spacing, often considered adequate for typical applications.

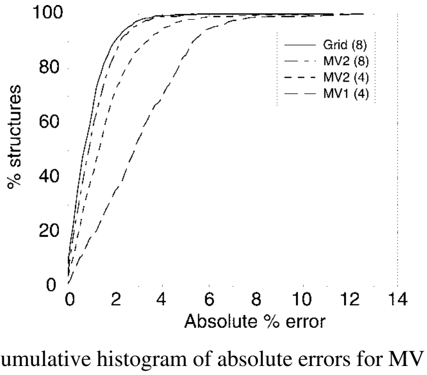

Figure 5. Cumulative histogram of absolute errors for MV1 and MV2 using a 611-protein structure data set. Y-value indicates percentage of structures that have better than X-value absolute percent error. Value in parentheses is K_s. Benchmark is SMV-based Poisson calculations performed with 0.25-Å grid spacing.

Table 2. Errors (kcal/mol) in Relative Energy Differences between all Pairs of 215 Conformations of Protein L.

	MV2 (4)	MV2 (8)	Grid (8)	MV1 (4)
Absolute avg.	10.04	9.24	9.56	19.72
SD	7.34	6.86	7.53	14.44
Max	42.34	43.05	47.84	79.94
Slope	0.980	0.988	0.980	0.918
Intercept	−1.738	−0.833	−1.882	−3.531
R	0.9944	0.9951	0.9946	0.9789

PARAM22 vdW radii and atomic charges are used. The value of K_s is denoted in parentheses. The slope and intercept of the best-fit line between GB energy pairs and Poisson energy pairs are also reported. *R* refers to the correlation coefficient. For reference, solvation energies for this data set range from about −1000 to −1400 kcal/mol.

The SASA-1 method that we introduced above was also evaluated. The benchmarks in this case are the exact calculations obtained using an analytic SASA module in CHARMM with a probe radius of 1.4 Å. In Figure 6, the SASA-1 method incurs errors for atomic surface areas on the order of ± 5 Å.[2] Table 4 indicates that errors on the order of 1–2% should be expected for SASA differences between pairs of conformations. For a protein like protein L, this error would amount to 1–2 kcal/mol assuming a hydrophobic energy term of 25 cal/(mol · Å^2). Also, as seen in Table 4, total surface areas are reliable for a span of small and large proteins, having a mean absolute error of less than 1%.

Finally, we would like to compare the computational times of the two analytic GB models that have been discussed in this work. In Table 5, one can see that the MV2 model is a bit slower than MV1. Also, the SASA-1 term is estimated to take a negligible amount of extra computational effort.

Discussion and Conclusions

The GB–MV2 model is a significant improvement in accuracy over the previous MV1 model of Lee et al. It is one of the most accurate analytic GB formalisms designed to mimic SMV-based Poisson solvation energies. However, comparisons with other GB methods should be performed to verify this conjecture.

It should be noted that our new technique is probably two to three times slower than most pair-based GB schemes.[2–4,28] Compared to Poisson method calculations of comparable accuracy, that is, 0.5-Å grid spacing, the GB–MV2 method is still quite a bit faster. For protein systems of about 100 amino acids, we find that the GB–MV2 procedure has about the same computational cost per timestep as an optimally prepared explicit solvent calculation, for example, one using a truncated octahedral volume with periodic boundary conditions and particle-mesh Ewald for treatment of long-range electrostatic interactions. Nonetheless, unlike explicit solvation calculations, GB has no friction arising from explicit water molecules and a great deal fewer degrees of freedoms. Thus, GB has much better efficiency per timestep for conformational sampling of the solute. Even so, there are several ways to speed up our method for molecular dynamics simulations: using a multiple timestep approach,[29] shortening the tail of the atomic function, and reducing the integration grid size.

The favorable performance of the GB–MV2 model further demonstrates that an analytic volume model that closely matches the SMV definition will provide accurate self-polarization energies and thus more accurate total solvation energies. The MV2 model is by no means identical to the SMV, but it is a sufficiently good approximation to obtain accurate self-polarization energies. It is

Table 3. Errors (kcal/mol) Versus SMV–Poisson (0.25 Å) for Relative Energy Differences of All Pairs of 120 Conformations of 1VII.

	MV2 (4)	MV2 (8)	Grid (8)	MV1 (4)
Absolute avg.	4.44	4.12	3.63	9.91
SD	3.46	3.32	2.77	7.48
Max	22.85	25.17	17.92	49.40
Slope	1.006	1.003	1.007	0.9902
Intercept	0.171	−1.175	0.393	3.988
R	0.9990	0.9992	0.9994	0.9957

PARAM22 vdW radii and atomic charges are used. The value of K_s is denoted in parentheses. The slope and intercept of the best-fit line between GB energy pairs and Poisson energy pairs are also reported. *R* refers to the correlation coefficient. For reference, solvation energies for this data set range from about −400 to −900 kcal/mol.

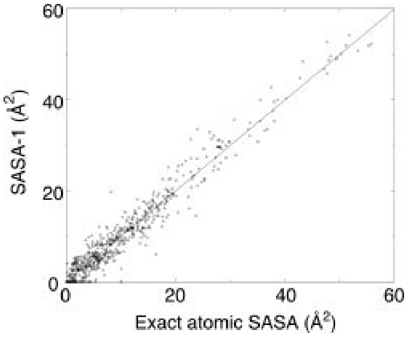

Figure 6. Scatter plots of atomic surface areas comparing the SASA-1 method versus the exact SASA calculation for protein L (2PTL). PARAM22 vdW radii are used. A y = x line is shown for comparison.

Table 4. SASA: Percentage Errors of SASA-1 Approximation vs. Exact for (a) Relative Differences of All Pairs of 215 Conformations of Protein L, (b) All Pairs of 120 Conformations for 1VII, and (c) a 611-Protein Test Set.

	Protein L	1VII	Protein set
Absolute avg.	1.21%	1.59%	0.82%
SD	0.96%	1.18%	0.68%
Max	6.97%	7.21%	3.53%

Formula for unsigned percentage error of a pair of structures $i,j = |100\% \cdot [(\sigma_i^{\text{SASA-1}} - \sigma_j^{\text{SASA-1}}) - (\sigma_i^{\text{exact}} - \sigma_j^{\text{exact}}) / \frac{1}{2}(\sigma_i^{\text{exact}} + \sigma_j^{\text{exact}})]|$.

Table 5. CPU Timing Results for 100 Steps of Minimization of the Protein 1AJJ[30] Using the PARAM22 Force Field and Infinite Nonbonded Cutoffs.

Solvent model	Time (s)	Relative time
Vacuum	4.04	1×
MV1	45.01	11.1×
MV2	56.80	14.1×
MV2+SASA-1	57.93	14.3×
SASA-1 only[a]	1.13	0.3×

[a] SASA-1 time is estimated. All timing runs were performed on a 250-MHz SGI Octane R10K.

definitely an improvement over a simple sum of atomic functions in the regimes where two spheres interpenetrate or are nearly touching. The one caveat so far is that the MV2 model introduces small artifacts at a certain range of interatomic distances.

The MV2 model is not limited to GB applications. Analytic Poisson models can adopt the MV2 volume definition to obtain better agreement with Poisson calculations using the well-established SMV. The MV2 model solves the important problem of removing gaps in the interior of proteins that would not be large enough to fit a water probe.

The analytic surface area method that we introduced in this article is similar to other pair-based methods in the literature. However, it is an exact method in the limit of an infinite number of angular integration points. For a finite number of integration points, SASA-1 yields adequate and adjustable accuracy for use as a hydrophobic/cavitation surface area energy term. It is also much faster than an exact surface area scheme when used in conjunction with the lookup table already generated for our MV1 and MV2 schemes.

In future publications we will assess the properties of the GB–MV2 model applied to general molecular dynamics simulations and to problems of specific interest such as ligand binding, pK_a calculations, and protein folding. We will also consider various modifications of the Still GB formula that may be used to obtain even better correspondence with Poisson calculations.

Acknowledgments

The authors thank Professors Alex Mackerell and Don Bashford, and Dr. Wonpil Im for their helpful discussions. Financial support from the National Institutes of Health to C.L.B. (GM37554) and F.R.S. (GM20334) is appreciated.

References

1. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. J Am Chem Soc 1990, 112, 6127.
2. Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. J Phys Chem 1996, 100, 19824.
3. Schaefer, M.; Karplus, M. J Phys Chem 1996, 100, 1578.
4. Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. J Phys Chem A 1997, 101, 3005.
5. Dominy, B. N.; Brooks, C. L., III. J Phys Chem B 1999, 103, 3765.
6. Scarsi, M.; Apostolakis, J.; Caflisch, A. J Phys Chem A 1997, 101, 8098.
7. Ghosh, A.; Rapp, C. S.; Friesner, R. A. J Phys Chem B 1998, 102, 10983.
8. Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. J Chem Phys 2002, 116, 10606.
9. Lee, B.; Richards, F. M. J Mol Biol 1971, 55, 379.
10. Richards, F. M. Annu Rev Biophys Eng 1977, 6, 151.
11. Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. J Phys Chem 1993, 97, 3591.
12. Im, W.; Beglov, D.; Roux, B. Comput Phys Commun 1998, 111, 59.
13. Grant, J. A.; Pickup, B. T.; Nicholls, A. J Comput Chem 2001, 22, 608.
14. Cortis, C. M.; Langlois, J. M.; Beachy, M. D.; Friesner, R. A. J Comput Chem 1996, 105, 5472.
15. Cossi, M.; Mennucci, B.; Cammi, R. J Comput Chem 1996, 17, 57.
16. Hasel, W.; Hendrickson, T. F.; Still, W. C. Tetrahedron Comp Met 1988, 1, 103.
17. Vasilyev, V.; Purisima, E. O. J Comput Chem 2002, 23, 737.
18. Jayaram, B.; Liu, Y.; Beveridge, D. L. J Chem Phys 1998, 109, 1465.
19. Salsbury, F. R., Jr.; Lee, M. S.; Feig, M.; Brooks, C. L., III. Ongoing work.
20. McQuarrie, D. A. Statistical Mechanics; University Science Books: Sausalito, CA, 2000.
21. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J Comput Chem 1983, 4, 187.
22. Abramowitz, M.; Stegun, I. A. Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables; Dover Publications: New York, 1974.
23. Nefcti, S. N. Introduction to the Mathematics of Financial Derivatives, 2nd ed.; Academic Press: San Diego, 2001.
24. Wodak, S. J.; Janin, J. Proc Natl Acad Sci USA 1980, 77, 1736.
25. MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, J. D.; Evanseck, M. J.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J Phys Chem B 1998, 102, 3586.
26. Wikstrom, M.; Drakenberg, T.; Forsen, S.; Sjobring, U.; Bjorck, L. Biochemistry 1994, 33, 14011.
27. McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. Nat Struct Biol 1997, 4, 180.
28. Dominy, B. N.; Brooks, C. L., III. CHARMM Version 28 Documentation; Harvard University: Boston, 1998.
29. Tuckerman, M.; Berne, B. J.; Martyna, G. J. J Chem Phys 1992, 97, 1990.
30. Fass, D.; Blacklow, S.; Kim, P. S.; Berger, J. M. Nature 1997, 388, 691.