A Residue-Pairwise Generalized Born Scheme Suitable for Protein Design Calculations

Georgios Archontis*,† and Thomas Simonson*,‡

Department of Physics, University of Cyprus, PO20537, CY1678, Nicosia, Cyprus, and Laboratoire de Biochimie (UMR7654 du CNRS), Department of Biology, Ecole Polytechnique, 91128 Palaiseau, France

Received: September 17, 2005

We describe an efficient generalized Born (GB) approximation for proteins, in which the interaction energy between two amino acids depends on the whole protein structure, but can be accurately computed from residue-pairwise information. Two results make the scheme pairwise. First, an accurate expression exists for the interaction energy between two residues R and R′ that depends on the product $B = B_R B_{R'}$ of their residue Born solvation radii. Second, this expression is accurately fitted by a parabolic function of B; the (three) fitting coefficients depend only on the pair RR′, *not* on its environment. In effect, the quantity B captures all the information that is relevant about the pair's dielectric environment. The method is tested with calculations on several hundred structures of the proteins trpcage, BPTI, ubiqutin, and thoredoxin. It yields solvation energies in better agreement with Poisson calculations than a traditional GB formulation. We also compute the effect of the protein/solvent environment on the interactions between pairs of charged residues in the active site of the enzyme aspartyl-tRNA synthetase. Our method captures this effect as accurately as traditional GB. Because it is residue-pairwise, the method can be incorporated into efficient protocols for rotamer placement and computational protein design.

Introduction

Computational protein design is an emerging technique that has been used to create new ligands, enzymes, and biosensors.[1−8] Since electrostatic interactions with aqueous solvent are crucial for protein stability,[9,10] accurate solvent models should be used. Continuum electrostatics provides the necessary accurate models.[11−13] Thus, generalized Born (GB) continuum electrostatic models are increasingly used to represent water implicitly in protein simulations.[14−22] The most recent variants yield an accuracy for structure and thermodynamics that is not much inferior to explicit solvent models.[19,22]

Computational protein design, however, has usually been done with simpler solvent models. Indeed, the space of possible protein sequences is gigantic. Computational design requires energy evaluations of many billions of protein sequences and side chain rotamer states, so that efficiency is paramount.[1−8] In particular, it is essential to express the protein solvation energy as a sum of residue-pairwise terms, each depending on a single amino acid pair. With a residue-pairwise energy, it is possible to precompute and store the residue−residue interaction energies for every residue pair, allowing for all possible combinations of side chain rotamers and amino acid types. If the protein is not too large, these data will not occupy more than a gigabyte of storage. The search for optimal sequences and structures can then be done very efficiently in a second step, using a lookup table that contains all the necessary energy information.[1−8] In contrast, if the energy is a many-body function, energies must be calculated on the fly while sequences and structures are searched, which is prohibitive for all but the smallest problems.

In continuum electrostatics, the effective interaction between two residues depends on the entire protein's shape and the complementary volume occupied by high dielectric solvent. Therefore, continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[13,23] Here, we overcome this difficulty and describe an accurate Generalized Born model that is residue-pairwise and can be used efficiently in protein design. Two essential results make the scheme pairwise. First, an accurate expression exists for the interaction energy between two residues R and R′ that depends on the product $B = B_R B_{R'}$ of their *residue* Born solvation radii. These radii (defined below) reflect the desolvation, or burial within the protein of each residue and are readily obtained from residue-pairwise quantities within GB models. Second, this expression is accurately fitted by a simple, parabolic function of B; the (three) fitting coefficients depend only on the pair RR′, *not* on its environment. In effect, the quantity B captures all the information that is relevant about the pair's dielectric environment.

By adopting an interaction energy that involves the residue Born solvation radii, we depart from current GB variants, which employ atomic Born solvation radii. In what follows, we will refer to the new variant as the "residue GB" approximation. Its accuracy is tested by calculations on five proteins. We consider several hundred conformations (each) of the proteins trpcage, BPTI, ubiquitin, and thioredoxin. In each case, the backbone has its experimental conformation, while side-chain conformations are randomized. The solvation energies of these structures are evaluated by the residue GB model, a standard, "atomic" GB method,[16] and a more accurate, continuum electrostatic model where Poisson's equation is solved numerically. In all cases, the residue GB gives a better agreement with the Poisson solvation energies than the atomic GB does. We also compute the electrostatic interactions between pairs of charged residues in the active site of aspartyl-tRNA synthetase. Each pair is held in a fixed, randomly chosen orientation while the conformations of the surrounding side chains are randomized. The residue GB

* Corresponding authors: archonti@ucy.ac.cy; thomas.simonson@polytechnique.fr.
† University of Cyprus.
‡ Ecole Polytechnique.

10.1021/jp055282+ CCC: $30.25
Published on Web 11/04/2005

model reproduces the variations of the pair interactions with an accuracy equivalent to the atomic GB model.

In the following theory section, we recall briefly the ingredients of standard, atomic GB. Residue GB is then obtained as a (slight) modification of an existing, atomic GB variant. Finally, we explain the idea leading to our fitting procedure and a GB that is fully pairwise at the residue level. In Numerical Methods, we describe the calculation of pair interaction energies and total solvation energies. This requires the choice of a force field and an initial atomic GB variant. We use the Charmm19 force field[24] along with GB/ACE.[16] GB/ACE has been used for peptide dynamics,[25] small molecule and protein solvation,[16,26] protein dynamics,[27] protein Xray structure refinement,[28] and protein acid/base titration.[29] Since residue GB is obtained as a modification of an existing atomic GB, we are primarily interested in the relative quality of the two. We expect that, while the choice of a different force field and/or atomic GB variant might lead to superior agreement with Poisson data or experiment,[29] the relative behavior of the atomic GB and its derived residue GB would be similar to the present GB/ACE case.

The Numerical Methods section also describes our choice of rotamer library, software to compute GB and Poisson energies, routines for parabolic fitting, and calculation and tabulation of residue solvation radii and interaction energies. The Results section presents data for five proteins of various sizes. The last section gives our conclusions.

2. Theory

2.1. Standard or Atomic GB. In GB models, the electrostatic energy includes both a direct, Coulomb term and a contribution from the solvent, polarized by the solute charges. Treating the solvent as a linear, homogeneous, dielectric medium, the total electrostatic energy has the form:

$$E^{\mathrm{elec}} = E^{\mathrm{Coul}} + \Delta G^{\mathrm{solv}} = \frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{r_{ij}} + \frac{1}{2}\sum_{ij} g_{ij} \tag{1}$$

where the sums are over all pairs of protein charges and the second sum includes diagonal terms, $i = j$. This second sum, ΔG^{solv}, represents the electrostatic solvation free energy of the protein (in the given conformation).[12] The term g_{ij} represents the interaction between a protein charge q_i and the solvent polarization induced by another charge, q_j. We refer to it as a GB interaction or screening energy. In the standard, atomic GB model,[14] this term is approximated by

$$g_{ij} = g(\underline{r}_i,\underline{r}_j) = \frac{\tau q_i q_j}{(r_{ij}^{\,2} + b_i b_j \exp[-r_{ij}^{\,2}/4b_i b_j])^{1/2}} \tag{2}$$

where $r_{ij} = |\underline{r}_i - \underline{r}_j|$, $\tau = 1/\epsilon_w - 1$, ϵ_w is the solvent dielectric constant (80 at room temperature), and b_i, b_j are effective, *atomic* "solvation radii" of the charges i, j.

The functional form of g_{ij} in eq 2 is the original and most frequently used form of Still et al.[14] The screening energy g_{ij} depends explicitly on the atomic positions $\underline{r}_i$ and $\underline{r}_j$ and implicitly on all the other atomic positions, through the solvation radii. Indeed, the radius b_i is determined by the "self" energy E_i^{self} of charge i:

$$E_i^{\mathrm{self}} = \frac{1}{2}g_{ii} \overset{\mathrm{def}}{=} \tau\frac{q_i^{\,2}}{2b_i} \tag{3}$$

E_i^{self} is the interaction energy between q_i and the polarization it creates in the solvent. In practice, b_i is roughly equal to the shortest distance between q_i and the protein surface. In the GB model, it is approximated by a simple, analytical function of the positions of all the solute atoms: $b_i = b_i(\underline{r}_1, \underline{r}_2, ..., \underline{r}_N)$. Different GB variants use different functional forms.[14−20] The essential point is that in most variants, including the ones considered here, the self-energy takes the form of a pairwise sum over atoms:

$$E_i^{\mathrm{self}} = \sum_j E_{ij}^{\mathrm{self}}(\underline{r}_i, \underline{r}_j) \tag{4}$$

The quantity E_{ij}^{self} depends only on $\underline{r}_i$, $\underline{r}_j$. It can be thought of as the free energy to replace solvent by the low dielectric solute in the volume of atom j when q_i is the only charge present (a more rigorous statement would require more discussion,[16] but is not needed here).

With the atomic radii defined by eq. 3, the total GB solvation free energy ΔG^{solv} has the correct behavior in the independent-atom limit (all atoms infinitely separated) and in the united-atom limit (all atoms coincide). This is a hallmark of GB models.[21]

2.2. Residue GB. It is straightforward to modify the above GB formulation to employ "residue" solvation radii, instead of the atomic radii b_i. This will lead to a "residue" GB. We define a self-energy contribution corresponding to a particular residue pair R,R′ by the expression

$$E_{RR'}^{\mathrm{self}} = \sum_{i\in R,\, j\in R'} E_{ij}^{\mathrm{self}} \tag{5}$$

where the double sum extends over atom pairs where i belongs to residue R and j to residue R′. Since the quantities E_{ij}^{self} depend only on the coordinates of the atomic pair ij (at least in the GB variants considered here), the self-energy $E_{RR'}^{\mathrm{self}}$ is a function only of the coordinates of the pair RR′. The self-energy of residue R can be written

$$E_R^{\mathrm{self}} = \sum_{R'} E_{RR'}^{\mathrm{self}} \tag{6}$$

and the total self-energy can be written

$$E^{\mathrm{self}} = \sum_R E_R^{\mathrm{self}} \tag{7}$$

This residue-decomposition of the self-energy is exact within the GB model, and follows immediately from the atom-pairwise character of eq 4. Analogous to eq 3, we then define the *residue* solvation radius B_{R} by the relation

$$E_{\mathrm{R}}^{\mathrm{self}} \overset{\mathrm{def}}{=} \tau\sum_{i\in R}\frac{q_{\mathrm{i}}^{\,2}}{2B_{\mathrm{R}}} \tag{8}$$

We also have

$$E_{\mathrm{R}}^{\mathrm{self}} = \sum_{i\in R} E_i^{\mathrm{self}} = \frac{\tau}{2}\sum_{i\in R}\frac{q_i^{\,2}}{b_i} \tag{9}$$

so that

$$\left(\sum_{i\in \mathrm{R}} q_i^{\,2}\right)\frac{1}{B_{\mathrm{R}}} = \sum_{i\in \mathrm{R}}\frac{q_i^{\,2}}{b_i} \tag{10}$$

Thus, B_R is a harmonic average over the b_i, $i \in R$, weighted by the squared charges.

We next define the contribution $g_{RR'}$ of residues R, R′ to the total screening energy ΔG^{solv}. In atomic GB, eq 2 gives

$$g_{RR'} = \sum_{i\in R,\, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + b_i b_j \exp[-r_{ij}^2/4b_i b_j])^{1/2}} \quad (11)$$

Here, we propose an alternative form:

$$g_{RR'} = \sum_{i\in R,\, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + B_R B_{R'} \exp[-r_{ij}^2/4B_R B_{R'}])^{1/2}} \quad (12)$$

For R = R′, the double summation in eqs 11−12 is actually restricted to distinct pairs of different atoms. Note that in both eq 11 and eq 12, $g_{RR'}$ depends on the entire protein sequence and structure (through B_R or b_i). Eqs. 11−12 have the same behavior at the independent-atom and united-atom limits. At intermediate distances, they constitute different interpolation schemes, whose relative accuracy can be tested numerically. In what follows, we refer to eq 12 as the "residue" GB approximation.

We notice that residue GB, eq 12, can be interpreted in another way. Starting from atomic GB, eq 11, it is easy to show mathematically that in situations of practical interest there exists a unique parameter $\mathscr{B}(R,R')$ such that

$$g_{RR'} = \sum_{i\in R,\, j\in R'} \frac{\tau q_i q_j}{(r_{ij}^2 + \mathscr{B}(R,R') \exp[-r_{ij}^2/4\mathscr{B}(R, R')])^{1/2}} \quad (13)$$

$\mathscr{B}(R, R')$ can be viewed as a nonlinear average over the products $b_i b_j$, $i \in R$, $j \in R'$, difficult to compute in practice. We obtain residue GB if *(i)* we continue to use the atomic GB interaction energy (eq 11), but (*ii*) we adopt $B_R B_{R'}$ as a heuristic guess for $\mathscr{B}(R,R')$. Although the previous interpretation of residue GB (as a new interpolation scheme) may seem simpler, the second is mathematically equivalent.

2.3 Making GB Pairwise: the Residue Self-Energy Is Sufficient to Parametrize the Pair Screening Energy. Residue GB is trivially pairwise for the self-energy and the solvation radii. However, it is not yet pairwise for the interaction, or screening energies. Indeed, the screening energy $g_{RR'} = g_{RR'}(B_R B_{R'})$ in (eq 12) explicitly depends on the solvation radii B_R, $B_{R'}$, which depend on the entire protein structure.

A fully pairwise scheme can be devised, however. We note that, for fixed interatomic distances r_{ij}, $g_{RR'}(B_R B_{R'})$ is a slowly varying function of $B_R B_{R'}$. This dependency can be approximated by a low-order polynomial:

$$g(B;r) = (r^2 + B \exp[-r^2/4B])^{-1/2} \approx c_1(r) + c_2(r)B + c_3(r)B^2 + O(B^3) \quad (14)$$

The interaction energy $g_{RR'}$ then takes the form

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 \quad (15)$$

For the five proteins studied below, this approximation holds for a large range of $B = B_R B_{R'}$ values. With eq 15, residue GB is fully residue-pairwise.

Indeed, consider a protein design problem with a given protein backbone, made up of $n = 100$ amino acids. We wish to compute the solvation energy corresponding to any set of amino acid types and any set of side-chain rotamers. The protein contains $n(n-1)/2$ amino acid pairs, each of which can have one of twenty types (e.g., alanine), and occupy on the order of 10 rotamers (depending on the exact choice of rotamer library). Thus, the total number of pair combinations, considering all positions, amino acid types, and rotamers, is on the order of $M = 1/2n(n-1) \times 20 \times 20 \times 10 \times 10 \approx 2 \times 10^8$. For each combination, the three corresponding fitting coefficients $c_i^{RR'}$ can be precomputed and stored (notice that they depend on the amino acid types and rotamers), along with the self-energy contributions $E_{RR'}^{self}$. Subsequently, for any given sequence and rotamer set, we can efficiently reconstruct the residue solvation radii B_R and the interaction energies $g_{RR'}$ from tabulated quantities, and the total solvation free energy can be obtained.

We show below that this residue-pairwise scheme is very accurate. We interpret this as follows: for a given pair R,R′ in a given environment (a particular protein sequence and set of side chain rotamers), the important information on the pair's dielectric environment is accurately captured by a single number: the solvation parameter $B_R B_{R'}$.

We note, finally, that in the proteins considered below, for a small percentage of the residue pairs, a more complicated fitting function is needed to obtain very high accuracy. The GB screening function has the following limiting behavior for large and small values of the radius parameter $B = B_R B_{R'}$:

$$(r^2 + B \exp[-r^2/4B])^{-1/2} \approx B^{-1/2} - \frac{3}{8}B^{-3/2} + O(B^{-5/2}), \quad B \gg r^2$$

$$(r^2 + B \exp[-r^2/4B])^{-1/2} \approx \frac{1}{r} - \frac{1}{2}\frac{y}{r^3}B + \frac{3}{8}\frac{y^2}{r^5}B^2, \; B \ll r^2 \quad (16)$$

where $y = \exp(-r^2/4B)$. Due to this limiting behavior, $g_{RR'}(B)$ can be approximated with excellent accuracy by a five-point function:

$$g_{RR'}(B) \approx c_1^{RR'} + c_2^{RR'}B + c_3^{RR'}B^2 + c_4^{RR'}B^{-1/2} + c_5^{RR'}B^{-3/2} \quad (17)$$

3. Numerical Methods

Energy Function. In what follows, the protein atomic charges and van der Waals radii were modeled by the CHARMM19 energy function.[24] The GB calculations were performed with the GB/ACE model,[16] implemented in XPLOR[30] as described in refs 26 and 28. The atomic volumes corresponded to the Voronoi database V01,[31] scaled by a factor of 0.8 as described in ref 27. For all calculations, the nonbonded interaction cutoff was infinite.

Generation of Random Protein Structures. For each of the proteins trpcage, BPTI, ubiquitin, and thioredoxin, random conformations were created by keeping the backbone fixed in its experimental conformation and positioning the side chains in random rotamers, taken from the library of Tuffery et al.[32] The side chains of Gly, Ala, Pro residues and Cys residues in disulfide bonds were kept fixed.

Calculation of Residue Solvation Radii. The residue solvation radii B_R were computed from a pretabulated self-energy matrix (see below) and eqs 6 and 7. Occasionally, the computed self-energy of a residue was small, yielding a large solvation radius B_R. In this case, we followed the method of Schaefer et al.:[33] a maximum radius B_{max} was defined, based on the protein

TABLE 1: Residue Interaction Energy Fitting to a Parabolic or a 5-Point Function

protein	total number of pairs	number requiring a 5-point fit	maximum χ^2 value[a]
trpcage	16410	727	1.2
BPTI	225602	5695	1.8
ubiquitin	509250	13538	2.0
thioredoxin	645994	11316	2.2

[a] Maximum total deviation (in (kcal/mol)^2) between the exact and fitted residue-pair energies using the 5-point function of eq 17, in the range B = 1−150 A^2. The quantity χ^2 is defined in eq 19.

TABLE 2: Standard Deviation (kcal/mol) between Poisson Energies (PE) and GB Energies

		standard deviation GB−PE	
protein	number of residues	atomic GB	residue GB
trpcage	20	21.2	13.6
BPTI	58	63.4	43.6
ubiquitin	76	97.3	43.9
thioredoxin	108	93.1	39.9

size; if $E_R^{self} > \tau\sum_{i\in R}q_i^2/2B_{max}$, the radius B_R was determined from the relation

$$B_R = 2B_{max}\left(1 - \frac{E_R^{self}B_{max}}{\tau\sum_{i\in R}q_i^2}\right) \quad (18)$$

The B_{max} values were 16 Å for trpcage, 25 Å for BPTI and ubiquitin, and 30 Å for thioredoxin.

Fitting Residue−Pair Energies to a Parabolic or 5-Point Function. The fitting of residue−pair interaction energies was performed with a Fortran program based on the general linear fit subroutine LFIT from Numerical Recipes.[34] The screening energies of all possible pairs were initially fit to a parabolic function of B in the range 1 to 150 Å^2. The quality of the fits was assessed by computing for each pair the merit function

$$\chi^2 = \sum_i (y_i^{RR'} - f^{RR'}(B_i))^2 \quad (19)$$

where $f^{RR'}(B)$ is the fitting function for pair RR′ (eq 15), the B_i are 20 evenly spaced values between 1 and 150 Å^2, and the $y_i^{RR'}$ are the "exact" residue−pair screening energies of the pair RR′, calculated by eq 12 for $B = B_i$. For pairs where χ^2 exceeded an (arbitrarily chosen) cutoff of 1 (kcal/mol)^2, the screening energies were fitted to the 5-point function of eq 17; the quality of the fits was assessed in the same way. The maximum value χ^2 for such (5-point) pairs did not exceed 2.2 (kcal/mol)^2 in all the cases studied here (see Results and Table 1).

Calculation of Residue−Pair Interaction Energies. For a particular protein structure, the GB interaction energy of each residue pair was computed from the pair coefficient $B_{RR'} = B_R B_{R'}$ and one of eqs 15 or 17. Occasionally, the value of $B_{RR'}$ was beyond the B range used in the fitting procedure: $B_{RR'} >$ 150 Å^2. In these cases, the screening energy was computed by a linear-order Taylor expansion of eqs 15 and 17 around the point $B_{RR'} = 150$ Å^2, using the corresponding first-order derivatives of eqs 15 and 17. This procedure was found to be more accurate than extending the B range beyond 150 Å^2 when deriving the fitting coefficients.

The Self-Energy Matrix. Residue pairs contribute to the self-energy through the terms $E_{RR'}^{self}$. Notice that both backbone and side chains contribute, and that diagonal terms R = R′ must be included. The $E_{RR'}^{self}$ are computed with the XPLOR program, allowing for all side chain rotamer combinations, and stored in a square matrix, which is nonsymmetric, since $E_{RR'}^{self} \neq E_{R'R}^{self}$ in general.[16] For concreteness, we describe the trpcage example in detail. Trpcage contains $n = 20$ amino acids, including 3 glycines, 4 prolines, and no alanines or cysteines. The remaining 13 amino acids can each occupy several rotamers (about 13 each, on average), determined by the Tuffery rotamer library.[32] Let k_i be the number of rotamers for side chain i. For the residue pair RR′ = (i, j), there are k_ik_j possible rotamer combinations. Summing over the $n(n-1)/2$ pairs gives a total of $\sum_{i,j}k_ik_j =$ 12740 distinct pair combinations for trpcage side chains. The total number of residue pair combinations for all proteins studied here is given in Table 1.

Aspartyl-tRNA Synthetase Calculations. The protein coordinates were taken from Protein Data Bank entry 1IL2. The Asp ligand was positioned in the catalytic site, using the known aspartyl−adenylate coordinates. Hydrogen atoms were positioned by the HBUILD facility of CHARMM. Protein and ligand were modeled by the CHARMM19 energy function. For five amino acid pairs (see the Results section) we generated 30−50 rotamer combinations each. Then, for each one, we created 500 random environments by altering the rotameric states of three different residues (excluding the considered pair) within a 10 Å sphere centered on the aspartic acid ligand. The atomic GB radii b_i were obtained by the ACE2 module of the CHARMM program.[24]

The standard deviation of the difference between the atomic GB and residue GB interaction energies for a particular combination i of rotamers of an amino acid pair RR′ (Table 3) was calculated by the equation

$$sd(RR', i) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x^2(RR', i, j) - \frac{1}{N^2}\left(\sum_{i=1}^{N}x(RR', i, j)\right)^2} \quad (20)$$

where sd(RR′, i) is the standard deviation for the ith rotamer combination of RR′, j labels the $N = 500$ environments, and x(RR′, i, j) is the difference between the atomic GB and residue GB interaction energies for the ith rotamer combination of RR′, surrounded by environment j. The mean standard deviation of a particular amino acid pair RR′, averaged over the corresponding rotamer combinations, was calculated by the equation

$$sd(RR') = \frac{1}{p}\sum_{i=1}^{p}sd(RR', i) \quad (21)$$

where p is the number of rotamer combinations considered for this pair (30−50 combinations).

Poisson Calculations. Poisson solvation energies were computed with the program UHBD.[35] The protein and solvent dielectric constants were set to 1 and 80, respectively. The protein−solvent dielectric boundaries were defined by the molecular surface of the protein. The atomic radii were taken from the CHARMM19 force field, with the exception of the hydrogen radii, which were set to 1.0.[36] The probe sphere for the molecular surface construction had a radius of 2 Å. Random positioning of interior side chains created occasional artificial voids in the protein interior. Even though these cavities were disconnected from bulk solvent, they could be assigned a high dielectric constant, even with the 2 Å probe radius. To prevent this, any such cavities were filled with dummy spheres. Permanent atomic charges were taken from the CHARMM19 force field. The Poisson equation was solved in two steps. The first step used a cubic grid spacing of 0.8 Å; the second, focusing step[37] used a spacing of 0.4 Å.

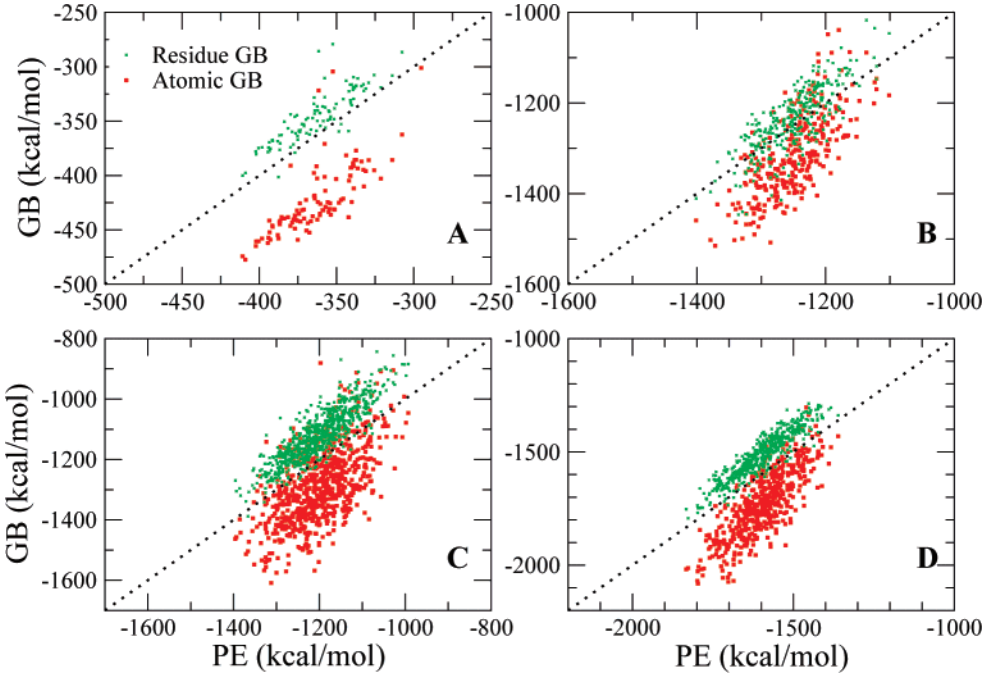

Figure 1. Residue and atomic GB solvation energies (kcal/mol) for several hundred random structures of trpcage (A), BPTI (B), ubiquitin (C), and thioredoxin (D), versus the corresponding Poisson energies.

TABLE 3: Standard Deviation (kcal/mol) between Atomic Residue GB Screening Energies for Five Residue Pairs in Aspartyl-tRNA Synthetase

amino acid pair	number of rotamer combinations	standard deviation	maximum standard deviation
K198-D233	49	0.17	0.70
E235-R489	31	0.24	0.79
E171-R217	47	0.46	0.54
K198-E235	42	0.11	0.55
K198−R489	41	0.17	0.70

4. Results

4.1. Total GB Energy Calculations. Atomic and residue GB are distinguished by their different interpolation schemes for the residue−residue interaction energy, eqs 11−12. To compare the accuracy of the two schemes, we calculated the total solvation energies with atomic GB, residue GB, and the Poisson equation for several hundred random structures of the proteins trpcage, BPTI, ubiquitin, and thioredoxin. For each protein, the computational procedure consisted of the following steps: (1) determine the self-energy matrix elements (eq 5) with XPLOR, allowing for all possible rotamer combinations, using appropriate XPLOR input files; (2) calculate the fitting coefficients $c_i^{RR'}$ (eq 15 or 17) for all residue pairs, allowing for all possible rotamer combinations; (3) generate random structures and compute the corresponding residue self-energies and solvation radii (eqs 6, 8); (4) compute the RR′ interaction energies from the pair coefficients $B_{RR'} = B_R B_{R'}$ and the appropriate fitting coefficients.

GB energies with the residue and atomic GB approximations are plotted in Figure 1, along with the corresponding energies from the Poisson equation (PE). The standard deviations of the PE−GB differences for all proteins are listed in Table 2.

In all cases, the residue GB method gives significantly better agreement with the PE. For the smallest protein (trpcage), the PE−GB rms deviation is 13.6 kcal/mol with residue GB and 21.2 kcal/mol with atomic GB. For the largest protein (thioredoxin), the corresponding rms deviations are 39.9 and 93.1 kcal/mol. Correlations between residue GB and PE are also superior. Since the self-energies of the two GB methods match by construction, the differences between atomic and residue GB come from the screening terms. These terms are positive in both cases, but the atomic GB values are smaller, giving total solvation energies that are too negative.

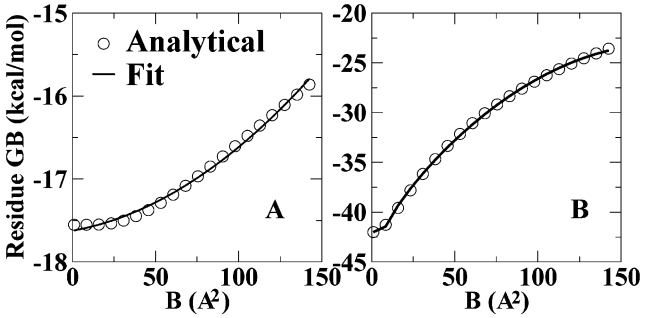

Figure 2. Fit of the residue GB interaction energy for the trpcage pair K8−R16 to a parabolic (A) or a five-point (B) function of $B = B_R B_{R'}$. (A) rotamer pair K8(rot17)-R16(rot29); (B) K8(rot23)-R16(rot14).

As described in the Theory section, the residue GB pair interaction energies $g_{RR'}$ can be fitted to a simple function of the parameter $B = B_R B_{R'}$. For each protein, the fitting was performed by an automated procedure described in the Numerical Methods section. Figure 2 shows two examples for two trpcage pairs. The fits are described, respectively, by the equations

$$g = -17.624648 + 0.003608911B + 0.000064803B^2$$

$$g = -40.143453 + 0.2023972B - 0.0005748B^2 - 9.485891B^{-1/2} + 7.424255B^{-3/2}$$

The total number of residue pairs, the number of pairs requiring a fit to a 5-point function, and the maximum observed deviation between the 5-point fit and the "exact" values (expressed by the merit function χ^2 of eq 19) are given in Table 1. For all proteins, the vast majority of residue pairs (96−98%) could be fitted to a parabolic function with a deviation $\chi^2 < 1$ (kcal/

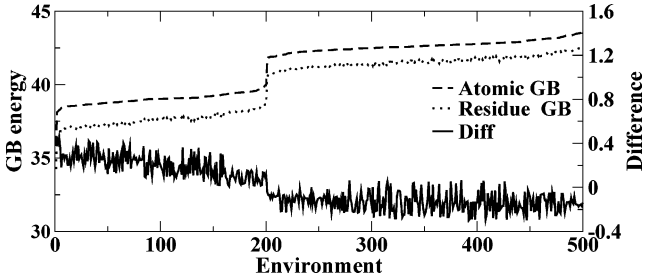

Figure 3. Atomic and residue GB interaction energies (kcal/mol) for the AspRS residue/rotamer pair K198(rot22)-E235(rot10) for 500 random protein environments, numbered by ascending energy. The difference between the atomic GB and residue GB values is shown (solid line and right *y* axis); the mean difference has been subtracted.

mol)^2. The few remaining pairs were fitted to the 5-point function; the fits obtained have a total deviation $\chi^2 < 1.2-2.2$ (kcal/mol)^2, depending on the protein.

4.2. Pair Energy Calculations for Aspartyl-tRNA Synthetase. The above calculations show that residue GB constitutes a satisfactory approximation for the total solvation energy. However, it is also important to reproduce the variations of individual pair energies in different environments. Such variations occur because the protein−solvent dielectric boundary changes when surface side chains change conformations. To test the behavior of residue GB, we chose five pairs RR′ of charged amino acids in the active site of the enzyme aspartyl-tRNA synthetase (AspRS). The pairs are listed in Table 3. These pairs play an important role in specific substrate recognition and have actually been targeted in a computational design study.[38]

For each of the five pairs RR′, we positioned the side chains of R and R′ in several different rotamers, generating 30−50 rotamer combinations. For each combination, we randomized the conformations of surrounding side chains within a 10 Å sphere, creating 500 random environments (per rotamer combination). The GB screening energies for a typical rotamer pair of the amino acids K198-E235 are shown in Figure 3.

We see that residue GB captures the variations of the screening energy as accurately as atomic GB. For the pair in Figure 3, the standard deviation of both the individual atomic GB and residue GB values is 1.8 kcal/mol, whereas the rms deviation (rmsd) between the two methods is just 0.19 kcal/mol. The rmsd between the atomic GB and residue GB values, averaged over rotamer pairs, varies between 0.1 and 0.5 kcal/mol (Table 3). The maximum rmsd for any residue pair is 0.8 kcal/mol, while the variability of the individual energies is significantly larger. Thus, for all these cases, the accuracies of residue GB and atomic GB are essentially equivalent.

5. Conclusions

Continuum electrostatic energies are many-body quantities that cannot ordinarily be expressed as a sum over residue or atom pairs.[13,23] Here, we have overcome this difficulty and described an accurate generalized Born model that is residue-pairwise. Equation 8 for the residue solvation radius, eq 12 for the residue interaction energy, and the fitting procedure (eq 15 or 17) provide the basis of the method. Because it is pairwise, residue GB is highly suitable for protein design calculations that search for amino acid sequences stabilizing a particular backbone fold or a particular protein−ligand complex. In such calculations, the protein can be subdivided into residues with a discrete set of allowed conformations (e.g., rotamers). The nonsymmetric self-energy matrix (eq 5) and the fitting coefficients $c_i^{RR'}$ for all possible pairs can be calculated in advance, allowing for all combinations of rotamers and amino acid types. This setup stage is performed only once and is moderately computationally expensive. For thioredoxin (108 amino acids), the self-energy matrix and the fitting coefficients required a few CPU hours on a single, 3.0 GHz Pentium processor. If full sequence variability is allowed for, the size of the calculation is increased by 20×20, giving a large, but still tractable (and trivially parallelizable) calculation, requiring little memory and about a gigabyte of storage.

After this setup stage is completed, the residue coefficients B_R can be constructed from this table and eq 8, for any amino acid sequence and any set of rotamers, and the screening energies can be evaluated by the corresponding fitting functions. In effect, for any pair RR′, the important information on the pair's current environment is captured by a single number: the current value of $B_R B_{R'}$.

From the data presented above, the accuracy of this scheme should be equivalent to the original, atomic GB model. The residue-pairwise scheme can be implemented in combination with any of several current or future GB variants.[22] As GB models continue to improve, the accuracy of our scheme should also improve. We[39] have performed extensive comparisons between GB/ACE and a simpler solvent model that is commonly used for protein design: a solvent accessible surface area model.[12] In contrast to some earlier work that used a nonoptimal GB/ACE parameterization,[40] GB does a good job of estimating the gain or loss of stability when charged amino acids are introduced or removed. Therefore, we expect that as GB variants improve, residue GB will perform increasingly better than several other solvent models usually employed for protein design,[1,2,4−8] including recent continuum electrostatic variants that achieve a pairwise form by neglecting long-range desolvation effects.[7,23] This will open exciting new possibilities for computational protein design.

Acknowledgment. Support was provided by the Egide program ZENON between Cyprus and France (to G.A. and T.S.).

References and Notes

(1) Koehl, P.; Levitt, M. *J. Mol. Biol.* **1999**, *293*, 1161.
(2) Kraemer−Pecore, C.; Wollacott, A.; Desjarlais, J. *Curr. Opin. Chem. Biol.* **2001**, *5*, 690−695.
(3) Wernisch, L.; Héry, S.; Wodak, S. *J. Mol. Biol.* **2000**, *301*, 713−736.
(4) Jaramillo, A.; Wodak, S.; Wernisch, L.; Héry, S. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *99*, 13554.
(5) Kuhlman, B.; Dantas, G.; Ireton, G.; Varani, G.; Stoddard, B.; Baker, D. *Science* **2003**, *302*, 1364−1368.
(6) Looger, L.; Dwyer, M.; Smith, J.; Hellinga, H. *Nature* **2003**, *423*, 185−190.
(7) Marshall, S.; Vizcarra, C.; Mayo, S. *Protein Sci.* **2005**, *14*, 1293−1304.
(8) Pokala, N.; Handel, T. *Protein Sci.* **2004**, *13*, 925−936.
(9) Schaefer, M.; Vlijmen, H. V.; Karplus, M. *Adv. Protein Chem.* **1998**, *51*, 1−57.
(10) Warshel, A.; Parson, W. *Q. Rev. Biophys.* **2001**, *34*, 563−679.
(11) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144−1149.
(12) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1−20.
(13) Simonson, T. *Rep. Prog. Phys.* **2003**, *66*, 737−787.
(14) Still, W. C.; Tempczyk, A.; Hawley, R.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127−6129.
(15) Hawkins, G.; Cramer, C.; Truhlar, D. *Chem. Phys. Lett.* **1995**, *246*, 122−129.
(16) Schaefer, M.; Karplus, M. *J. Phys. Chem.* **1996**, *100*, 1578−1599.
(17) Qiu, D.; Shenkin, P.; Hollinger, F.; Still, W. *J. Phys. Chem. A* **1997**, *101*, 3005−3014.
(18) Ghosh, A.; Rapp, C.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(19) Lee, M.; Salsbury, F., Jr.; Brooks, C., III *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(20) Onufriev, A.; Case, D.; Bashford, D. *J. Comp. Chem.* **2002**, *23*, 1297−1304.
(21) Bashford, D.; Case, D. *Ann. Rev. Phys. Chem.* **2000**, *51*, 129−152.
(22) Feig, M.; Brooks, C. L., III *Curr. Opin. Struct. Biol.* **2004**, *14*, 217−224.
(23) Wisz, M.; Hellinga, H. *Proteins* **2003**, *51*, 360−377.
(24) Brooks, B.; Bruccoleri, R.; Olafson, B.; States, D.; Swaminathan, S.; Karplus, M. *J. Comp. Chem.* **1983**, *4*, 187−217.
(25) Schaefer, M.; Sommer, M.; Karplus, M. *J. Phys. Chem. B* **1998**, *101*, 1663−1683.
(26) Wagner, F.; Simonson, T. *J. Comp. Chem.* **1999**, *20*, 322−335.
(27) Calimet, N.; Schaefer, M.; Simonson, T. *Proteins* **2001**, *45*, 144−158.
(28) Moulinier, L.; Case, D.; Simonson, T. *Acta Crystallogr. D* **2003**, *59*, 2094−2103.
(29) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167−4180.
(30) Brünger, A. T. *X-PLOR version 3.1, A System for X-ray crystallography and NMR*; Yale University Press: New Haven, 1992.
(31) Schaefer, M.; Bartels, C.; Leclerc, F.; Karplus, M. *J. Comp. Chem.* **2001**, *22*, 1857−1879.
(32) Tuffery, P.; Etchebest, C.; Hazout, S.; Lavery, R. *J. Biomol. Struct. Dynam.* **1991**, *8*, 1267.
(33) Schaefer, M.; Bartels, C.; Karplus, M. *J. Mol. Biol.* **1998**, *284*, 835−847.
(34) Press, W.; Flannery, B.; Teukolsky, S.; Vetterling, W. *Numerical Recipes*; Cambridge University Press: Cambridge, 1986.
(35) Madura, J.; Briggs, J.; Wade, R.; Davis, M.; Luty, B.; Ilin, A.; Antosiewicz, J.; Gilson, M.; Baheri, B.; Scott, L.; Mccammon, J. *Comp. Phys. Comm.* **1995**, *91*, 57−95.
(36) Mohan, V.; Davis, M.; Mccammon, J. A.; Pettitt, B. M. *J. Phys. Chem.* **1992**, *96*, 6428−6431.
(37) Gilson, M.; Sharp, K.; Honig, B. *J. Comp. Chem.* **1988**, *9*, 327−335.
(38) Archontis, G.; Simonson, T.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 307−327.
(39) Lopes, A.; Archontis, G.; Simonson, T., manuscript in preparation.
(40) Jaramillo, A.; Wodak, S. *Biophys. J.* **2005**, *88*, 156−171.