An Improved Pairwise Decomposable Finite-Difference Poisson–Boltzmann Method for Computational Protein Design

CHRISTINA L. VIZCARRA,[1] NAIGONG ZHANG,[2] SHANNON A. MARSHALL,[1] NED S. WINGREEN,[3] CHEN ZENG,[2] STEPHEN L. MAYO[1,4]

[1]*Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125*

[2]*Department of Physics, George Washington University, Washington, DC 20052*

[3]*Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544*

[4]*Division of Biology, California Institute of Technology, Pasadena, California 91125*

Received 16 December 2006; Revised 1 October 2007; Accepted 17 October 2007

DOI 10.1002/jcc.20878

Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: Our goal is to develop accurate electrostatic models that can be implemented in current computational protein design protocols. To this end, we improve upon a previously reported pairwise decomposable, finite difference Poisson–Boltzmann (FDPB) model for protein design (Marshall et al., Protein Sci 2005, 14, 1293). The improvement involves placing generic sidechains at positions with unknown amino acid identity and explicitly capturing two-body perturbations to the dielectric environment. We compare the original and improved FDPB methods to standard FDPB calculations in which the dielectric environment is completely determined by protein atoms. The generic sidechain approach yields a two to threefold increase in accuracy per residue or residue pair over the original pairwise FDPB implementation, with no additional computational cost. Distance dependent dielectric and solvent-exclusion models were also compared with standard FDPB energies. The accuracy of the new pairwise FDPB method is shown to be superior to these models, even after reparameterization of the solvent-exclusion model.

 J Comput Chem 00: 000–000, 2007

Key words: Poisson–Boltzmann; protein design; continuum solvation; electrostatics

Introduction

Current computational protein design programs could be improved by the inclusion of an accurate model for electrostatics. Since proteins exist in highly polarizable solvents, the accuracy of the electrostatics model is dependent on the accuracy of the solvation model. To overcome the computational demands of explicitly modeling all of the water molecules in a macromolecular system, a continuum dielectric description of water is used in many biomolecular applications.[1,2] In continuum solvation models, the protein is treated as a low dielectric cavity within a high dielectric solvent. The boundary between the protein and solvent dielectric is defined by the protein's molecular surface. When carrying out amino acid sequence selection for protein design, the location of the dielectric boundary becomes ambiguous because the final amino acid identities and their conformations are not known until the very end of the calculation. In order to overcome this limitation and to satisfy the need for computationally efficient energy functions, alterations to the Generalized Born model,[3,4] a modified version of the Tanford–Kirkword model,[5,6] and various empirical models[7–10] have been reported for protein sequence design.

Within the limitations of a continuum solvent description, the Finite Difference Poisson–Boltzmann (FDPB) model is often considered a standard for accuracy.[11,12] A general strategy for implementing an FDPB model that is pairwise decomposable by sidechain conformation (rotamer) has been reported.[13] This strategy involves evaluating explicit perturbations to the dielectric boundary. For example, the desolvation energy of a sidechain on being transferred from the unfolded state to the folded state is calculated by solving for the difference in solvation energy between the one-body state (i.e., the folded backbone and one sidechain) and the unfolded state model for the sidechain. Two-

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192-8651/suppmat

Correspondence to: S. L. Mayo; e-mail: steve@mayo.caltech.edu

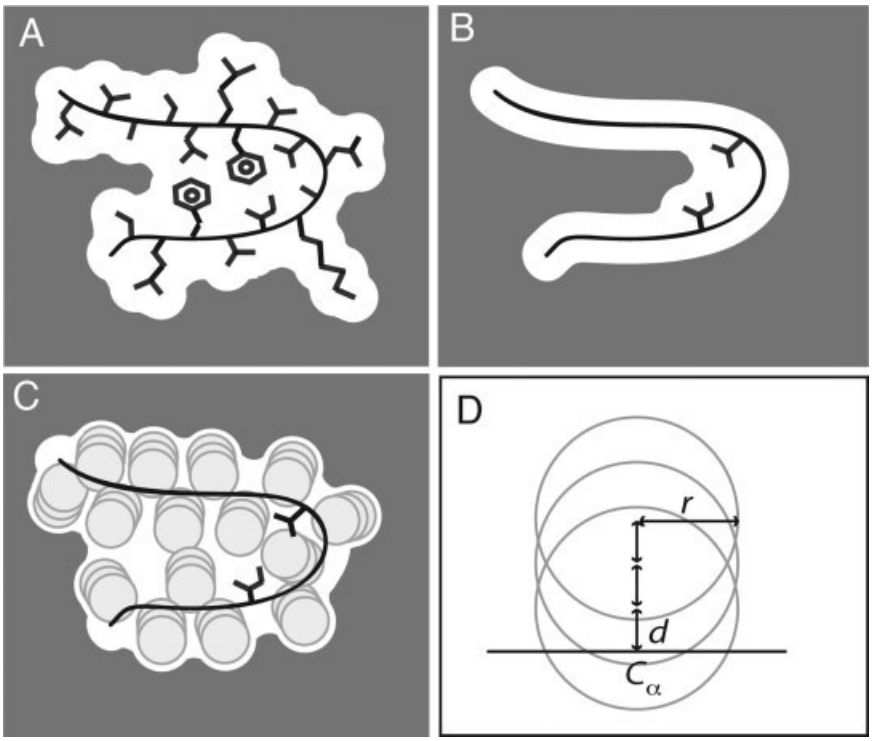

Figure 1. Illustration of exact, no generic sidechain (G0), and generic sidechain (G3) calculations. The dark gray area denotes solvent dielectric ($\varepsilon = 80$) and the white area denotes protein dielectric ($\varepsilon = 4$). (A) The exact molecular surface is defined by the backbone and all sidechains of the protein. (B) The two-body state for the G0 model is defined by two sidechains and the protein backbone. (C) The two-body state for the G3 model is defined by the two sidechains, the protein backbone, and three-sphere generic sidechains at all other positions. The one-body state is analogous to (B) or (C) but with only one sidechain represented explicitly. (D) The definition of radius and distance for the three-sphere generic sidechain.

body perturbations are calculated as the difference in solvation energy between a state with two sidechains (the ''two-body state'' in Fig. 1B) and the one-body state. The total pairwise sidechain desolvation is thus the desolvation of the sidechain by the backbone plus the sum of two-body perturbations. The energy terms in this method are fully pairwise decomposable by sidechain conformation and are therefore compatible with the energy matrices and optimization algorithms used in most computational design methods.

The accuracy of the pairwise decomposable FDPB model was assessed by comparing the energy calculated with the entire molecular surface defined by all of the protein sidechains (the ''exact'' surface in Fig. 1A) to the energy calculated using the sum of perturbations method. It was found that the desolvation of sidechains could be accurately approximated with an RMS error of 0.64 kcal mol^{-1} per sidechain.[1] The generic sidechains described by Zhang et al.[14] for calculation of pairwise solvent-accessible surface area present a straightforward and efficient strategy for improving the accuracy of pairwise approximate FDPB calculations. Figure 1 shows the difference between the original pairwise FDPB model and the generic sidechain approach. At all positions for which the identity or conformation of the amino acid is unknown, a generic sidechain composed of three spheres is placed, making the one-body state more closely resemble the true protein molecular surface and the two-body perturbations less dramatic.

A generic sidechain approach to approximating the volume occupied by a protein's sidechains has been used previously in many applications, including residue classification with respect to the molecular surface,[15,16] protein–protein docking,[17] and solvation.[14] Pokala and Handel have reported a one-body generic sidechain formulation of the Generalized Born (GB) model.[3] For each residue in a design calculation, they approximate the low dielectric environment by spheres at all other positions. We take a similar approach using the FDPB model, but, importantly, we also calculate two-body perturbations that lead to a better approximation of the protein environment. In order to overcome the computational limitations of an $O(n^2)$ calculation, distance cutoffs are tested.

Because of the computational demands of solving the PB equation numerically, there is a great deal of interest in methods that approximate the PB model, such as the GB model, and also in fast empirical models.[1] The solvent-exclusion model of Lazaridis and Karplus (LK)[18] is computationally efficient and has

been used by Baker and coworkers in the successful design of a novel fold.[19] Here we test the improved pairwise FDPB model against the LK model. Since the original parameterization of the LK model was based primarily on experimental solvation free energies, we derive new parameters based on FDPB energies to see how well the functional form of the LK model is able to reproduce this particular benchmark. While it is found that the generic sidechain method out-performs the LK model, the trade-off between computational efficiency and accuracy of the energy function is discussed.

Methods

FDPB Calculations

A set of 24 proteins with hydrogens added was taken from the Richardson Top 500 database of high resolution X-ray crystal structures (http://kinemage.biochem.duke.edu/databases/top500.php). The PDB codes for the set are: 1IGD, 1MSI, 1KP6, 1OPD, 1FNA, 1MOL, 2ACY, 1ERV, 1DHN, 1WHI, 3CHY, 1ELK, 2RN2, 1HKA, 3LZM, 1AMM, 1XNB, 153L, 1BK7, 2PTH, 1THV, 1BS9, 1AGJ, and 2BAA. A subset of 10 structures was used in the generic sidechain parameter optimization: 1IGD, 1KP6, 1FNA, 2ACY, 1DHN, 3CHY, 2RN2, 3LZM, 1BK7, and 1THV. The DelPhi program[20] was used to solve the linearized Poisson–Boltzmann equation using the following settings: 2 grids $Å^{-1}$, 0.05M salt, a protein dielectric of 4 and a solvent dielectric of 80. In all calculations on a single structure, the protein's position relative to the grid was held constant. PARSE radii and charges were used.[21] The test set contains 2028 polar residues when using PARSE charge definitions. All prolines and disulphide-bonds were treated as part of the backbone.

The three-sphere generic sidechain method (herein referred to as G3) reported by Zhang et al.[14] was used in all calculations described below unless otherwise noted. Calculations denoted G0 refer to the method of Marshall et al.[13] A grid-based search was carried out to find a more optimal set of generic sidechain dimensions for FDPB calculations. Within the grid search, the parameters reported previously,[14] sphere radius = 2.85 Å and distance between spheres = 0.61 Å, were found to be near-optimal and were used as given. A detailed description of the parameter search is given in the supplementary information.

To be consistent with the ORBIT force field,[22] the following terms were computed: (i) *backbone desolvation*, the energetic cost of desolvating the protein backbone by sidechains; (ii) *sidechain desolvation*, the energetic cost of polar sidechains being desolvated by the backbone and by other sidechains; (iii) *sidechain/backbone screened Coulombic energy*, the solvent screened electrostatic interaction energy between sidechains and the backbone, and (iv) *sidechain/sidechain screened Coulombic energy*, the solvent screened electrostatic interaction energy between pairs of sidechains. Each of these terms was computed with all protein atoms present in their crystallographic position (''exact'') and using reduced representations of the protein (''pairwise''). The pairwise calculations are analogous to those described in Marshall et al.[13] However, for all G3 calculations, three-sphere generic sidechains were used at all positions for which no sidechain was present. The unfolded state reference for sidechain desolvation consisted of the sidechain i plus the local backbone atoms: $C_\alpha(i-1)$, $C(i-1)$, $O(i-1)$, $N(i)$, $H(i)$, $C_\alpha(i)$, $C(i)$, $O(i)$, $N(i+1)$, $H(i+1)$, and $C_\alpha(i+1)$. Screened Coulombic interactions were only calculated in the folded state.

The most notable difference between the G0 and G3 methods is the calculation of the backbone desolvation energy (ΔG^{bb}_{desolv}). The unfolded state for the backbone is still described by a crystallographic backbone with no sidechains present. A ''zero-body'' folded state is then defined by the backbone with generic sidechains at all positions. Single residue (one-body) perturbations to the zero-body state are summed to get the total one-body backbone desolvation energy:

$$\Delta G^{bb}_{desolv} = \Delta G^{bb}_{zero-body} - \Delta G^{bb}_{unfolded} + \sum_{i}^{n} \left(\Delta G^{bb,i}_{one-body} - \Delta G^{bb}_{zero-body}\right) \quad (1)$$

By equation 1, the backbone desolvation energy is derived from $(n+2)$ DelPhi calculations, where n is the number of residues in the protein. In order to calculate all of the one- and two-body energies for a structure with n residues, p of which are polar, a total of

$$(n+2) + 2p + p(n-p) + p(p-1) \quad (2)$$

DelPhi calculations are needed, where $(n+2)$ corresponds to backbone desolvation, $2p$ corresponds to the unfolded state and one-body folded state models, $p(n-p)$ corresponds to perturbations of polar residues by nonpolar residues, and $p(p-1)$ corresponds to perturbations and interactions between polar residues, noting that two-body perturbations are not symmetric.

Distance Dependent Dielectric and Lazaridis–Karplus Calculations

New parameters for the solvent-exclusion model originally reported by Lazaridis and Karplus (LK)[18] were derived using the following 14 structure training set: 1MSI, 1OPD, 1MOL, 1ERV, 1WHI, 1ELK, 1HKA, 1AMM, 1XNB, 153L, 2PTH, 1BS9, 1AGJ, and 2BAA. While the CHARMM19 parameters described by LK[18] are atom-based, the new parameters are sidechain-based. For instance, lysine was assigned a single parameter for all heavy atoms with non-zero partial atomic charges in the PARSE charge set: C_ε and N_ζ. Since the desolvation energy of a sidechain in the LK model is independent of the ΔG_{ref} parameter, only values for ΔG_{free} were derived by fitting to the ''exact'' FDPB desolvation energy

$$\Delta G^{i}_{desolv} = \Delta G^{i}_{free} \sum_{t \in i} \sum_{u \notin i, local\ bb} f_t(r_{tu}) V_u \quad (3)$$

where the function f is the Gaussian free energy density of atom t and V_u is the volume of desolvating atom u. For each sidechain i in the training set, the sum of Gaussian solvent exclusion terms was calculated over each atom t in sidechain i and each atom u

Table 1. Accuracy of the Electrostatic Models.

	G0[a]		G3[b]	
	RMSD (kcal mol^{-1})	R	RMSD (kcal mol^{-1})	R
Backbone desolvation energy				
One-body	3.96	0.997	3.51	0.998
Sidechain desolvation energy				
One-body	1.93	0.718	0.79	0.915
Two-body, all pairs	0.64	0.962	0.40	0.979
Two-body, pairs <6 Å	0.67	0.968	0.35	0.984
Two-body, pairs <4 Å	0.82	0.952	0.39	0.980
Sidechain/backbone screened Coulombic energy				
One-body	0.90	0.957	0.34	0.987
Two-body, all pairs	0.36	0.987	0.18	0.996
Two-body, pairs <6 Å	0.41	0.984	0.17	0.996
Two-body, pairs <4 Å	0.51	0.979	0.23	0.994
Sidechain/sidechain screened Coulombic energy				
Two-body, all pairs	0.13	0.948	0.05	0.987
Two-body, pairs <6 Å[c]	0.13	0.939	0.06	0.979
Two-body, pairs <4 Å[d]	0.13	0.933	0.07	0.972

[a]Marshall et al.[13]
[b]Sphere radius = 2.85 Å, distance from C_α and distance between spheres = 0.61 Å.
[c]For pairs separated by more than 6 Å, a distance dependent dielectric constant of $4.93r$ was used.
[d]For pairs separated by more than 4 Å, a distance dependent dielectric constant of $5.56r$ was used.

that is not in the sidechain i or its local backbone. For each set of amino acids, with the following amino acids considered together: Asn and Gln; Ser and Thr; and Asp and Glu, a linear least squares fit was used to get ΔG_{free} from the exact FDPB sidechain desolvation energy ΔG_{dsolv}. The new LK parameters are listed in Table S1. For the distance dependent dielectric (DDD) calculations, dielectrics were assigned as $5.1r$ for sidechain/backbone interactions and $7.1r$ for sidechain/sidechain interactions, according to Zollars et al.[10] These values were derived by fitting to FDPB screened Coulombic energies. Since there were five structures in common between their training set and the 24 structures used here, the error in screened Coulombic energy was assessed for the remaining 19 structures.

Results

The accuracy of the pairwise FDPB methods was measured by comparison to ''exact'' FDPB energies calculated with all protein atoms present. The RMS error between the exact energies and the G0 and G3 models are listed in Table 1. The pairwise energies from the GO and G3 models are plotted against the exact energies in Figures 2 and 3. The G3 model performs better than the G0 model in all cases. As expected, the one-body sidechain desolvation improves dramatically over the G0 model in which only desolvation of the sidechain by the backbone is counted (Figs. 2C and 2D). Similarly, the one-body G3 model is more effective than the one-body G0 model at capturing the descreening of strong sidechain/backbone interactions (Figs. 2E and 2F). It is interesting to note that the one-body G3 model for sidechain desolvation is more accurate than the two-body G0 model with a 4 Å cutoff (Table 1). This indicates that the approximate surface provided by the generic sidechains is more effective at reproducing the exact energies than adding the one-body energy to the truncated sum of two-body sidechain perturbations in the G0 model, a relevant model to consider since a distance cutoff will almost certainly be used in design calculations.

As shown in Table 1 and Figure 3, the G3 two-body models for sidechain desolvation and sidechain/backbone screened Coulombic energy provide improvements over the one-body models and the G0 two-body models. An especially dramatic improvement in accuracy is seen for the two-body approximation for sidechain/sidechain interactions. Each data point in Figures 3E and 3F corresponds to a pair of residues. For the G0 model, there are no descreening contributions from other sidechains to sidechain/sidechain interactions, whereas the generic sidechains provide a vast improvement by approximating the reduced dielectric of other sidechains. The accuracy of the model is only slightly reduced by using an approximate distance dependent dielectric model for pairs separated by more than the specified cutoffs. The dielectric values were based on those reported by Marshall et al.[13] For both cutoff values tested, more than 90% of all polar sidechain pairs in the test set were not treated with an FDPB calculation. Such cutoffs would provide a considerable speed enhancement in the energy calculation stage of a design calculation.

Although the G3 model performs better than the G0 model for sidechain desolvation, there are noticeable outliers in Figure 3B. There was also one residue in the test set with a negative G3 two-body desolvation energy which is not shown in Figure 3B but is included in the calculated error in Table 1. Out of the 2028 residues in the test set, there are 19 for which sidechain desolvation is underestimated by more than 1.5 kcal mol^{-1} when using the G3 model. For this set of 19 residues, the amino acid types are exclusively Asp, Glu, Arg, and Lys,

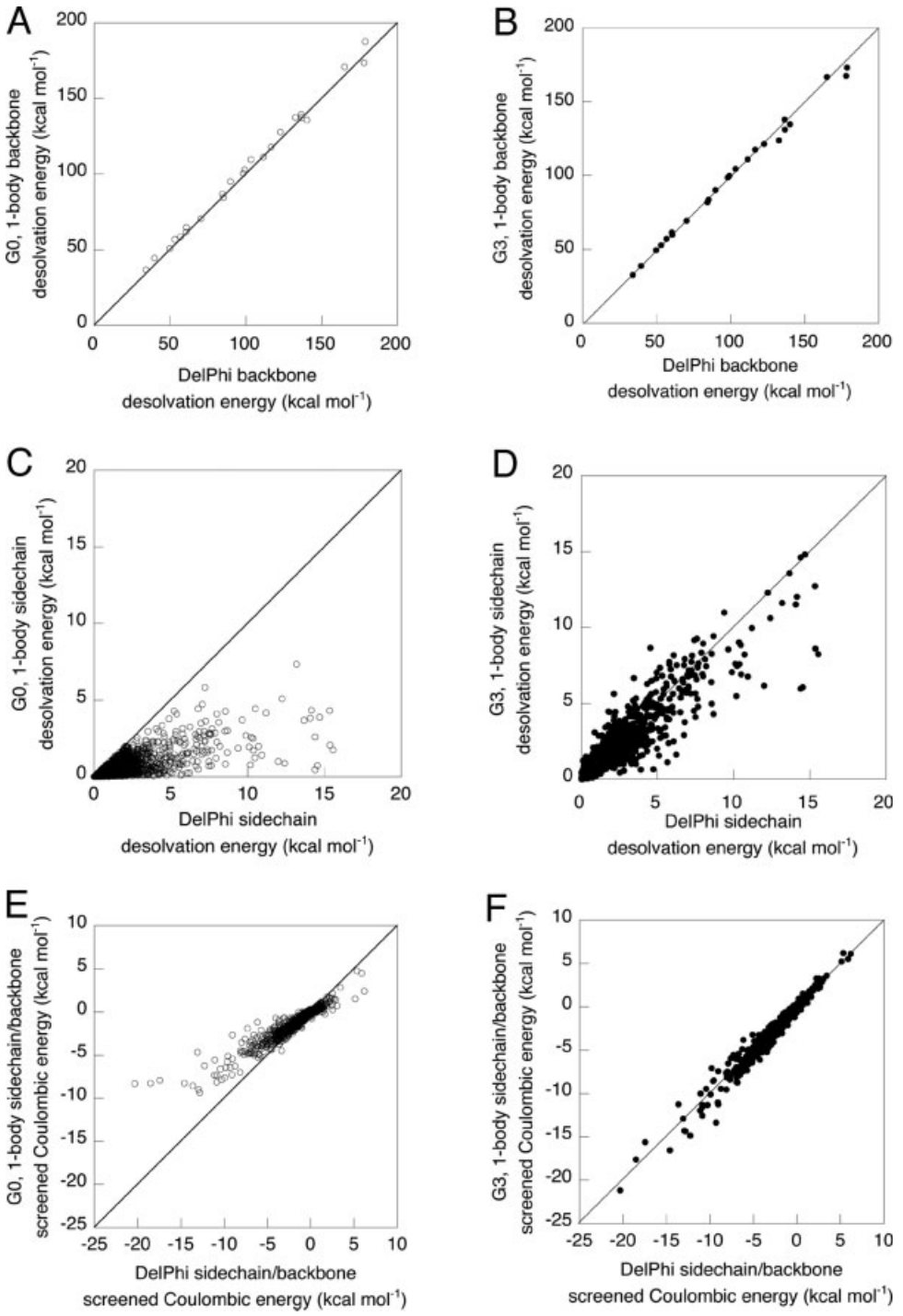

Figure 2. Accuracy of one-body G0 and G3 FDPB methods. One-body backbone desolvation calculated using the (A) G0 and (B) G3 methods. One-body sidechain desolvation calculated using the (C) G0 and (D) G3 methods. One-body screened Coulombic interaction energy between sidechains and backbone calculated using the (E) G0 and (F) G3 methods.

and they are from 12 different structures. Two of these outliers, shown as ''X'' symbols in Figure 3B, plus the point with a negative desolvation were sensitive to moving the molecule slightly with respect to the grid. This sensitivity to grid placement has been discussed previously for the pairwise FDPB calculation[13] and for more standard applications.[23] Although it would

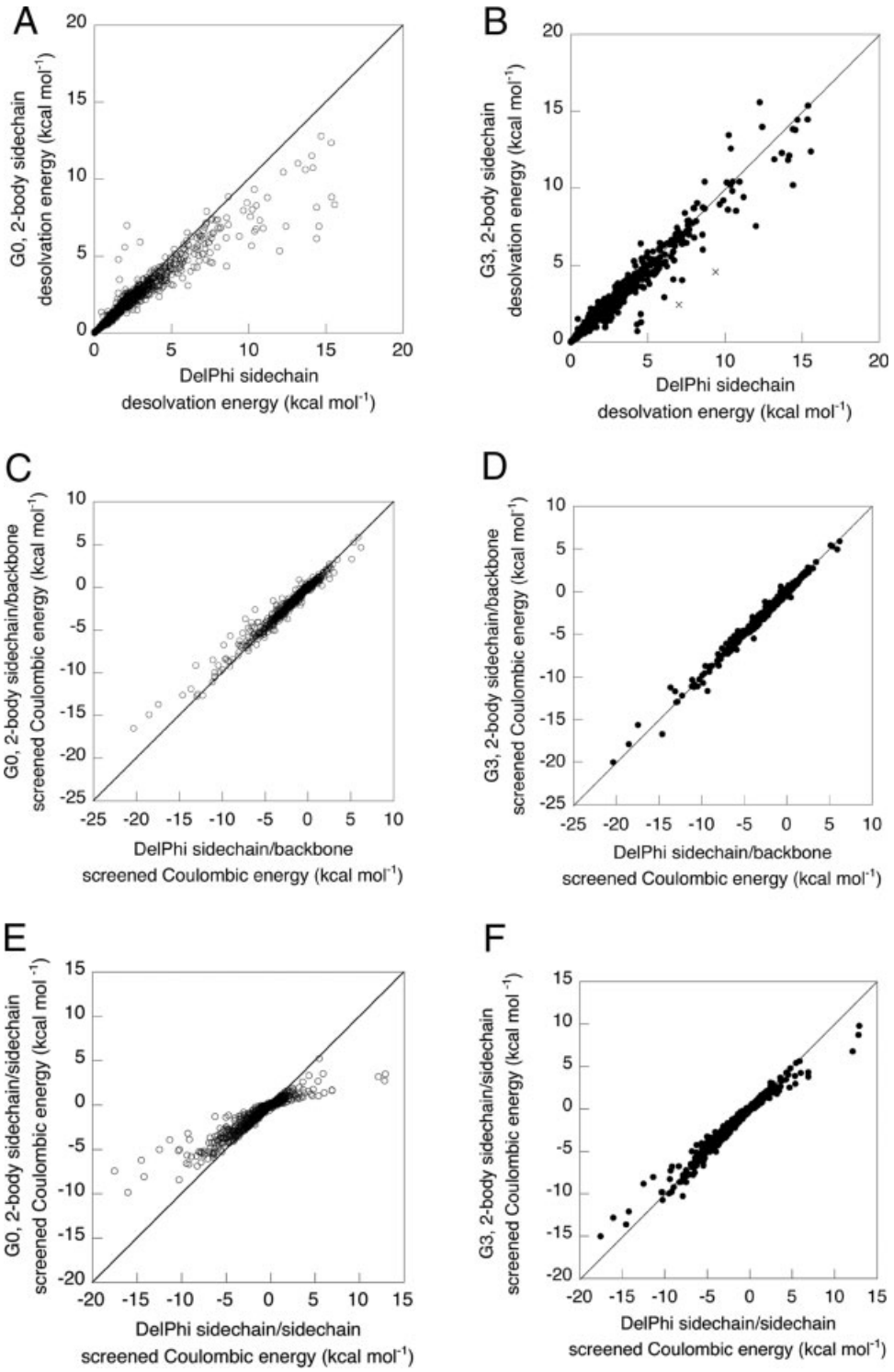

Figure 3. Accuracy of two-body G0 and G3 FDPB methods. Two-body sidechain desolvation calculated using the (A) G0 and (B) G3 methods. Points marked with ''X'' in (B) correspond to sidechains for which the desolvation energy is sensitive to the placement of the protein with respect to the grid. Two-body screened Coulombic interaction energy between sidechains and backbone calculated using the (C) G0 and (D) G3 methods. Two-body screened Coulombic interaction energy between pairs of sidechains calculated using the (E) G0 and (F) G3 methods.

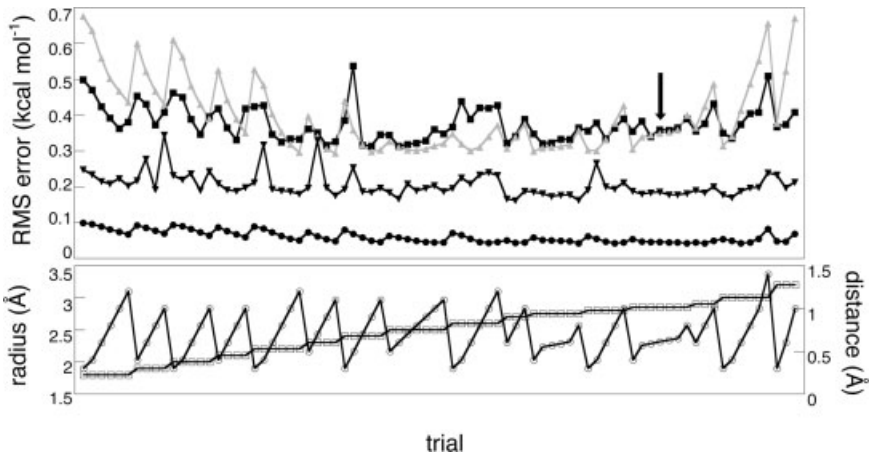

Figure 4. Sensitivity of the G3 FDPB method to generic sidechain parameters. Each line shows the error in a different force field component: two-body sidechain desolvation (■), one-body sidechain/backbone screened Coulombic energy (▲), two-body sidechain/backbone screened Coulombic energy (▲), and two-body sidechain/sidechain screened Coulombic energy (●). The lower panel shows the radius (□) and distance (○) that were sampled in each trial. The parameter set radius = 2.85 Å and distance = 0.61 Å is indicated by an arrow.

increase the calculation time *n*-fold, averaging over *n* translations with respect to the grid would alleviate this problem. An additional seven of the outliers had nearby residues that gave large, negative perturbations, leading to two-body energies with larger error than the one-body approximation. In five of the seven cases, these large negative perturbations are caused by glycines, the amino acid for which the G3 approximation is the most inaccurate. The remaining 10 cases with large, negative error had exact desolvation energies greater than 8 kcal mol^{-1} and one-body error greater than the two-body error, indicating that these points are simply difficult to capture by a pairwise summation scheme. For both two-body sidechain desolvation and sidechain/backbone screened Coulombic energy, the error decreases slightly when cutoffs are imposed. This may point to the fact that only local perturbations are necessary with the G3 model and that inclusion of longer-range perturbations leads to errors in accounting for sidechain overlap, as described by Zhang et al.[14]

The generic sidechain parameters used here are the same as those used in Zhang et al.[14] for solvent accessible surface area calculations. Since an alternative set of parameters may be more optimal for the molecular surface definition used in FDPB calculations, we carried out a grid search of parameters to find a superior set of parameters and to assess how sensitive the G3 method is to generic sidechain dimensions. As shown in Figures 4 and S1, the error is sensitive to sphere size and spacing over the entire parameter space explored, but parameter sets near radius = 2.85 Å and distance = 0.61 Å have relatively low error. The optimality of these parameters for both surface area and FDPB calculations supports the assertion that generic sidechains of these dimensions accurately represent the average space occupied by amino acid sidechains in folded proteins.[24] Therefore, if the surface definition (e.g., solvent-accessible or molecular surface) is consistent between the pairwise and exact calculations, these parameters will be suitable. It is also notable that the training set with which these parameters were originally

Table 2. Parameter Sensitivity of the G3 Model.[a]

Grid spacing (grids Å^{-1})	2	2	2	4[b]
Ionic strength (mM)	50	150	50	50
Translations No.	1	1	3	1
One-body backbone desolvation	3.51	3.57	3.30	3.13
One-body sidechain desolvation	0.79	0.81	0.79	0.80
Two-body sidechain desolvation[c]	0.40	0.41	0.38	0.41
One-body sidechain/backbone screened Coulombic energy	0.34	0.34	0.34	0.34
Two-body sidechain/backbone screened Coulombic energy[c]	0.18	0.18	0.19	0.18
Two-body sidechain/sidechain screened Coulombic energy[c]	0.05	0.06	0.05	0.06

[a]Error is reported as RMSD in kcal mol^{-1}.
[b]Only the exact calculations were carried out with a grid spacing of 4 grids Å^{-1}, while the one- and two-body calculations were carried with a grid spacing of 2 grids Å^{-1}.
[c]All two-body calculations were carried out without distance cutoffs.

Table 3. Comparison of FDPB, LK, and DDD Models.

	RMSD (kcal mol^{-1})	R
Sidechain desolvation energy[a,b]		
Two-body G0	0.53	0.969
Two-body G3	0.36	0.979
LK (CHARMM19)[c]	1.90	0.897
LK (tuned)	0.73	0.914
Sidechain/backbone screened Coulombic energy[a,d]		
Two-body G0	0.37	0.986
Two-body G3	0.19	0.996
DDD, $\varepsilon = 5.1r$[e]	0.83	0.921
Sidechain/sidechain screened Coulombic energy[a,d]		
Two-body G0	0.13	0.943
Two-body G3	0.05	0.986
DDD, $\varepsilon = 7.1r$[e]	0.14	0.915

[a]All two-body calculations were carried out without distance cutoffs.
[b]Ten structures in test set: 1IGD,1KP6, 1FNA, 2ACY, 1DHN, 3CHY, 2RN2, 3LZM, 1BK7, 1THV.
[c]Lazaridis and Karplus.[18]
[d]Ninteen structures in test set: 1MSI, 1KP6, 1OPD, 1FNA, 1MOL, 2ACY, 1ERV, 1DHN, 3CHY, 1ELK, 1HKA, 1XNB, 153L, 1BK7, 2PTH, 1THV, 1BS9, 1AGJ, 2BAA.
[e]Zollars et al.[10]

derived has an overlap of only one protein with the 10 structures used in the parameter search here, suggesting that these parameters are robust for general protein design targets.

Depending on the computational demands of a particular modeling application, different FDPB parameters may be preferable.[11] In order to assess the sensitivity of the error in the G3 model to parameter changes, we have sampled several possible FDPB parameter sets that might be used in standard calculations. As shown in Table 2, any changes in error values with ionic strength and translation averaging are small. We also compared the approximate numbers obtained using the G3 model at a grid spacing that would be reasonable for a design calculation (2 grids Å^{-1}) with an exact calculation carried out at finer grid spacing (4 grids Å^{-1}). The right-hand column in Table 2 shows the RMS error between these energies.

It is of general interest to see how the pairwise approximate FDPB method described here performs in comparison to fast pairwise decomposable methods already used for protein design.[1] We compared the performance of the solvent-exclusion model of LK[18] and a DDD model with the G3 method. Both of the LK and DDD models are highly parameterized. Since multiple parameter sets have been reported for the LK model, we tried both the CHARMM19 LK parameters[18] and a new set of parameters tuned specifically to reproduce PB energies. We used the distance dependent dielectric values that led to a stabilized designed protein in a recent experimental protein design study.[10] The results of this comparison are shown in Table 3 and Figures 5 and 6. The G3 model is more effective than the LK and DDD models at approximating exact FDPB calculations. While the performance of the LK model (Figs. 5B and 5C) varies greatly with parameters, the LK model has a nonlinear relationship with exact PB desolvation energies regardless of parameter set.

Discussion

We have shown that it is possible to improve the agreement between a pairwise decomposable FDPB method and an exact FDPB method with no additional computational cost. The improvement stems from the more accurate approximation of the dielectric boundary provided by generic sidechains. The RMS errors for screened Coulombic interactions between polar sidechains and between polar sidechains and the protein backbone are decreased by nearly threefold and twofold, respectively. The error associated with the desolvation of polar sidechains is reduced by nearly twofold. For the two-body perturbation-based

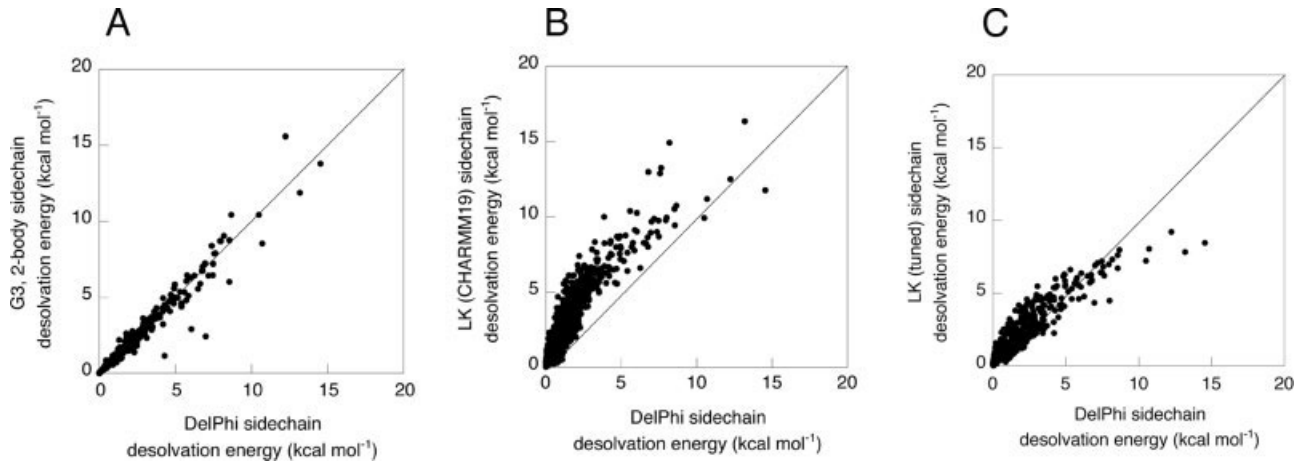

Figure 5. Accuracy of the G3 model (A) versus the LK solvent exclusion model (B,C) for approximating sidechain desolvation. Results for both the CHARMM19 (B) and tuned (C) LK parameter sets are shown. All plots contain data for the 758 polar sidechains from the 10 structures listed in Table 3 and described in the methods section.

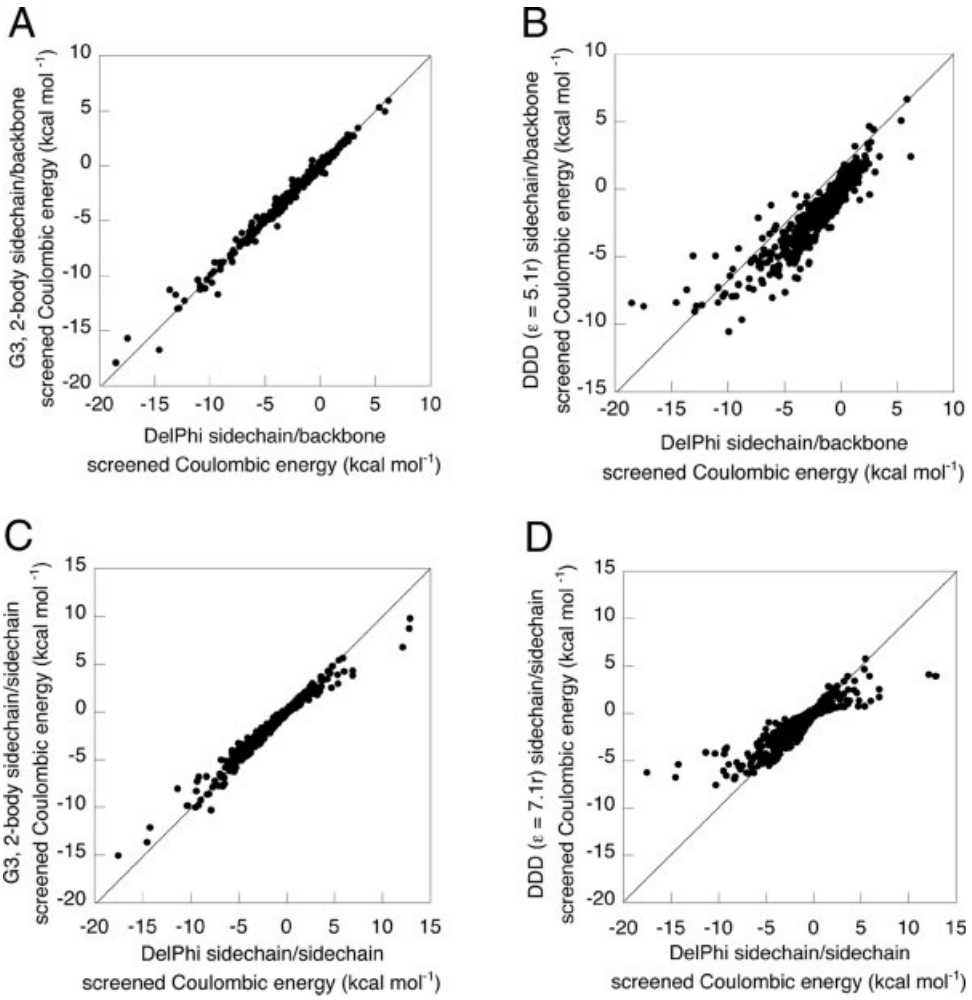

Figure 6. Accuracy of the G3 model (A,C) versus the DDD model (B,D) for approximating sidechain/backbone and sidechain/sidechain screened Coulombic interactions. Data is shown for the 19 structures listed in Table 3 and described in the Methods section.

terms (i.e., sidechain desolvation and sidechain/backbone screened Coulombic energy), this more accurate description of the dielectric boundary leads to less dramatic perturbations to that boundary, accounting for the inherent nonadditivity of such perturbations more effectively than the previously reported pairwise FDPB method.[13]

Ideally, a protein design energy function is both accurate and computationally efficient. The FDPB methods are three to four orders of magnitude slower than the standard ORBIT energy function. For example, a surface design of the small 51-residue helical protein engrailed homeodomain gave 6.4 million rotamer pairs for 29 design positions. The precalculation of all rotamer singles and pairs energies required on the order of ~9 CPU hours using the standard ORBIT energy function with a surface area based solvation term, ~0.1 CPU hours using a modified version of the ORBIT energy function with a DDD term and the LK model, and ~1000 CPU hours using the G3 model. This large computational cost requires one to carefully assess the appropriateness of the G3 model for different design problems. Because of the large investment of time and resources involved in synthesizing and characterizing designed proteins, an expensive calculation may be worthwhile, especially if the calculation involves positions with important electrostatic contacts such as in the active site of an enzyme. Unfortunately, the cost of the FDPB models may preclude large design targets since the calculation also scales poorly with the size of the grid on which the PB equation is solved.

The results shown here indicate good agreement between standard many-body electrostatic energies and those derived from summing one- and two-body perturbations. In a protein design calculation, the one- and two-body perturbations are stored in a look-up table, which is used to find the optimal sequence. Standard FDPB calculations serve as a reasonable benchmark for assessing the sequence energies that will be evaluated by summing one- and two-body perturbations from the table of rotameric energies. For search algorithms such as Monte

Carlo[25] or FASTER,[26] where total sequence energy is evaluated and used as a criteria to accept or reject rotamer changes, this benchmark is sufficient. However, the comparison with standard FDPB calculations leaves the possibility of a cancellation of error when summing over perturbations: some perturbations may be ''too small'' and some may be ''too large.'' This may become important when using algorithms based on Dead End Elimination.[27] In such algorithms, the sequence energy is not evaluated, but instead two-body energies are used to eliminate rotamers that are not in the optimal sequence. There is no clear way to gauge the accuracy of the individual perturbation energies when comparing to standard benchmarks since there is no ''exact'' perturbation energy. Indeed, the most stringent test of this improved electrostatics term will be in the context of a protein design force field and, ultimately, in experimental validation of designed protein sequences.

References

1. Koehl, P. Curr Opin Struct Biol 2006, 16, 142.
2. Vizcarra, C. L.; Mayo, S. L. Curr Opin Chem Biol 2005, 9, 622.
3. Pokala, N.; Handel, T. M. Protein Sci 2004, 13, 925.
4. Archontis, G.; Simonson, T. J Phys Chem B 2005, 109, 22667.
5. Havranek, J. J.; Harbury, P. B. Proc Natl Acad Sci USA 1999, 96, 11145.
6. Havranek, J. J.; Harbury, P. B. Nat Struct Biol 2003, 10, 45.
7. Marshall, S. A.; Morgan, C. S.; Mayo, S. L. J Mol Biol 2002, 316, 189.
8. Wisz, M. S.; Hellinga, H. W. Proteins 2003, 51, 360.
9. Cerutti, D. S.; Jain, T.; McCammon, J. A. Pro Sci 2006, 15, 1579.
10. Zollars, E. S.; Marshall, S. A.; Mayo, S. L. Protein Sci 2006, 15, 2014.
11. Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. J Comput Chem 2004, 25, 265.
12. Baker, N. A. Curr Opin Struct Biol 2005, 15, 137.
13. Marshall, S. A.; Vizcarra, C. L.; Mayo, S. L. Protein Sci 2005, 14, 1293.
14. Zhang, N. G.; Zeng, C.; Wingreen, N. S. Proteins 2004, 57, 565.
15. Dahiyat, B. I.; Mayo, S. L. Science 1997, 278, 82.
16. Marshall, S. A.; Mayo, S. L. J Mol Biol 2001, 305, 619.
17. Huang, P. S.; Love, J. J.; Mayo, S. L. J Comput Chem 2005, 26, 1222.
18. Lazaridis, T.; Karplus, M. Proteins 1999, 35, 133.
19. Kuhlman, B.; Dantas, G.; Ireton, G. C.; Varani, G.; Stoddard, B. L.; Baker, D. Science 2003, 302, 1364.
20. Rocchia, W.; Alexov, E.; Honig, B. J Phys Chem B 2001, 105, 6507.
21. Sitkoff, D.; Sharp, K.; Honig, B. J Phys Chem 1994, 98, 1978.
22. Gordon, D. B.; Marshall, S. A.; Mayo, S. L. Curr Opin Struct Biol 1999, 9, 509.
23. Gilson, M. K.; Sharp, K.; Honig, B. J Comput Chem 1987, 9, 327.
24. Creighton, T. E. Proteins: Structure and Molecular Properties; W.H. Freeman: New York, 1993.
25. Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. J Chem Phys 1953, 21, 1087.
26. Desmet, J.; Spriet, J.; Lasters, I. Proteins 2002, 48, 31.
27. Desmet, J.; De Maeyer, M.; Hazes, B.; Lasters, I. Nature 1992, 356, 539.