JCTC Journal of Chemical Theory and Computation

Improved Methods for Side Chain and Loop Predictions via the Protein Local Optimization Program: Variable Dielectric Model for Implicitly Improving the Treatment of Polarization Effects

Kai Zhu, Michael R. Shirts, and Richard A. Friesner*

Department of Chemistry, Columbia University, New York, New York 10027

Received July 3, 2007

Abstract: This paper presents significant improvements in both accuracy and computational efficiency of protein side chain and loop predictions using the Protein Local Optimization Program (PLOP). We introduce a novel energy model in which the internal dielectric constant of the protein is allowed to vary as a function of the interacting residues and present a physical rationale for this model. Using this model, we achieve qualitative improvements in the accuracy of side chain predictions with respect to experimental crystal structure and substantially reduce the RMSDs for loop predictions, particularly those predictions involving charged side chains. For the single side chain prediction of lysine, 40% of the errors are eliminated, and the accuracy increases from 62.6% to 76.8%. The errors in glutamate and aspartate predictions are reduced by 19% and 24%, respectively. When applied to a set of 240 loop predictions with 6, 8, 10, and 13 residue of loop length, this new model yields unprecedented accuracies with average backbone root-mean-square deviations of 0.39 Å, 0.68 Å, 0.80 Å, and 1.00 Å for 6, 8, 10, and 13 residue loops, respectively. We also describe a series of technical improvements in the PLOP simulation algorithms, which lead to a speedup of a factor of 2−4 in loop predictions.

I. Introduction

In several previous publications, we have described a novel approach to high-resolution protein structure prediction, implemented in the protein local optimization program (PLOP).[1−4] This program combines sophisticated conformational sampling algorithms with a molecular mechanics force field[5,6] and a continuum solvation model based on the generalized Born approach,[7,8] in contrast to typical programs for loop and side chain modeling which rely on either simplified physical chemistry based scoring functions or knowledge based potentials inspired by bioinformatics approaches.[9−21] While such approximate methods have performed respectably for low-resolution protein modeling, it appears as though achievement of a truly accurate atomic level description of protein structure—as is required for many practical applications, for example structure based drug design—necessitates the use of more accurate energy functions and correspondingly efficient and precise sampling algorithms. Using this more physical approach, we have achieved significant reductions in root-mean-square-deviation (RMSD) from crystal structures in repredictions of loops and side chains within the context of the native structure of the protein. Particularly large advances[4] are apparent for long loops (up to ∼11−13 residues), which place severe demands upon the accuracy of the scoring function and the efficiency of the sampling algorithms.

Despite these successes, the previously published methodology still displays systematic errors for subsets of test cases such as the prediction of lysine side chain structures. It also suffers from substantial computational requirements, which becomes a significant problem when applying the methods to problems such as homology modeling where the "context" of the local region to be refined is imperfect, and a large number of calculations per loop region are presumably

* Corresponding author e-mail: rich@chem.columbia.edu. Corresponding author address: Department of Chemistry, Columbia University, 3000 Broadway, Mail Code 3110, New York, NY 10027.

10.1021/ct700166f CCC: $37.00
Published on Web 10/16/2007

required to achieve convergence of the energy function. These deficiencies have motivated further development of both the computational algorithms in PLOP and the energy model used to rank order structures. Our expectation is that such improvements will be ongoing for a number of years, and our experience to date has been that both quantitative and qualitative advances in the technology continue to be generated by these efforts.

This paper presents significant improvements in both accuracy and computational efficiency. Qualitative improvements which both increase the accuracy of side chain prediction and substantially reduce the RMSDs for loop prediction are achieved by a model in which the internal dielectric constant of the protein is allowed to vary as a function of the interacting residues. A novel physical rationale for this model, based on an analysis of the treatment of polarization in contemporary fixed charge force fields, is presented, and the model is shown to have a particularly large effect on predictions for charged side chains. Second, we describe a series of technical improvements in the PLOP simulation algorithms, which lead to a speedup of a factor of 2−4 in loop prediction. Further acceleration of the calculations is clearly possible but is left for another publication.

The paper is organized as follows. In section II, a brief review of the PLOP methodology for side chain and loop prediction is presented. Section III introduces the variable internal dielectric model and provides the physical interpretation of this model as well as detailing its particular implementation in the current version of PLOP. It also presents the algorithmic improvements in the current version of PLOP responsible for increased speed. Section IV discusses the data sets that we use to benchmark the methodology, for both side chain and loop prediction. Section V presents accuracy and timing results, comparing earlier versions of PLOP as well as some literature data with the current version. Finally, in the Conclusion, we summarize the results and discuss future directions.

II. Side Chain and Loop Prediction via PLOP

We have described our side chain and loop prediction algorithms in previous publications.[1−4] Here we give only a brief review. Initially, we use *single* side chain prediction (i.e., keeping the remainder of the protein fixed at the native configuration) to parametrize and validate the variable dielectric model. We originally developed this strategy to develop and test the new torsional parameters for the OPLS-AA force fields[3] and for a novel protonation state assignment algorithm.[22] We use a hierarchical approach to single side chain prediction. Initially, side chain conformations are sampled using a highly detailed rotamer library developed by Xiang and Honig.[15] This library contains, for example, 2086 rotamers for lysine. The use of such a detailed library ensures adequate sampling. The associated computational expense is reduced by prescreening the rotamers using only hard sphere overlap as a criterion, which can be made very rapid with the use of a cell list. Many rotamers can in this manner be excluded before performing energy evaluations. Then a rapid, reduced energy calculation is performed for each remaining rotamer. The reduced energy uses a short cutoff for nonbonded interactions and includes only the torsional energy among the bonded terms. Next we perform a clustering procedure on these rotamers. We start at the lowest energy structure and then find all neighbors in torsional space, working outward until the energy no longer goes up. This is the first energy basin (or cluster). Then we find the lowest energy structure among the remaining rotamers and continue to do this for all energy basins until we run out of rotamers. The representative of each cluster is chosen as the lowest energy structure of the energy basin. The entropic contribution is calculated by taking the configurational integral in the torsional space over each basin. Then the representative is completely energy minimized using a fast minimization algorithm previously developed.[23] The sum of the minimized energy and the entropic contribution is used to rank the structures and give the final prediction.

Loop predictions in PLOP also feature a hierarchical approach. The generation of loop conformations is accomplished via a dihedral angle buildup procedure which, at the limit of highest resolution, exhaustively searches the phase space of possible loop geometries connecting the two loop stems. The energy evaluation achieves both efficiency and high accuracy via deployment of a hierarchy of scoring functions; rapid screening functions are used to eliminate large numbers of high-energy loops, ultimately yielding a relatively small number of candidates which are then clustered. Representative members of each cluster are then evaluated via minimization of an all atom molecular mechanics energy function and continuum solvation model (in this study OPLS-AA force field[5,6] plus SGB/NP solvent model).[7,8] Furthermore, we have developed a powerful sampling algorithm for the long loop predictions, which involves multiple stage loop predictions and refinements, and achieved very high accuracy when combined with a hydrophobic term we have developed to fix a major flaw in the generalized Born model.[4] The crystal environment is explicitly included in our loop and side prediction algorithms by using dimensions and the space group reported in PDB files.[1,2] PLOP executables can be obtained from Matthew Jacobson at UCSF, free of charge for academic users, as per instructions on his Web site (http://francisco.compbio.ucsf.edu/~jacobson/). An implementation of PLOP, with a graphical user interface, is also available to both academic and commercial users in the Prime program, distributed by Schrodinger, Inc.

III. Methods

A. Variable Internal Dielectric Model. The question of what value to use for the internal dielectric constant in Poisson−Boltzmann (PB) and generalized Born (GB) calculations has been the subject of a large number of papers over the past 20 years. Much of the early work was focused on PB methods, at a time when analytical gradients for PB calculations did not exist, and there was therefore no convenient way to carry out accurate conformational searches, geometry optimizations, or molecular dynamics simulations of PB based models. In this situation, movement of protein

groups was often invoked to justify the use of a "high" internal dielectric constant, typically, in the 4−20 range. For example, such values were used extensively in PB based pK_a calculations.[24−28] However, our current methodology involves a more extensive exploration of conformational phase space, so this component of the protein dielectric should not be contributory.

A second alternative is to assign an internal dielectric to the protein based on the optical dielectric constant, which is based on electronic response, not nuclear motion, of a "typical" organic compound. This value is on the order of 2.[29,30] At first glance, this appears to be a satisfactory approach, as there is no question that reorganization of the electrons in the protein will occur in response to an applied field. However, when considering what internal dielectric to employ, one has to take into account how the molecular mechanics (MM) force fields used in PB or GB calculations were developed. Even hydrocarbons in the OPLS-AA force field, for example, have small, but noticeable, point charges associated with them. The critical point is that in general the charges used in any MM force field are *already* enhanced from gas-phase values, to take into account the "average" polarization in the condensed phase. For example, the OPLS-AA force field was parametrized to fit experimental thermodynamic data such as density and heat of vaporization of pure liquids. An alternative is to use quantum chemically derived charges that are "scaled up", either by using a relatively small basis set that implicitly yields higher charges or by explicit use of a scale factor. Thus, one could argue that the optical polarization has been implicitly included in the force field, and using a dielectric of 2 results in double counting of this effect. One reasonable alternative is to employ specially derived charges that are fit to experiment to complement the model with a dielectric of 2, as in the PARSE model of Honig and co-workers.[31] However, this involves a complete redesign of the force field and discards the substantial amount of information obtained from fitting molecular dynamics simulations of pure liquids to thermodynamic data.

Our philosophy to date within PLOP has been to use a dielectric constant of 1, on the theory that optical polarization of the protein is incorporated into the force field as discussed above. However, there is one situation where this argument fails, and that is when a protein group is interacting with a charged species. Neutral groups, such as the hydroxyl group in serine, are parametrized to fit pure liquid simulations of methanol; hence, the assumed environment of the group is neutral hydrogen bonds to other hydroxyls. If a serine is instead hydrogen bonded to a lysine, the polarization of the group is presumably greater—but this is not reflected in the uniform internal dielectric of 1 that is used in the GB model in PLOP. Similarly, charged groups are parametrized to agree with solvation free energies of the charged species in water, again placing the group in question in a neutral environment. When a salt bridge is formed, the internal dielectric of 1 is then, again, presumably inappropriate.

The clearest solution to this problem is to employ a polarizable force field in high-resolution protein simulations. This is a promising future path that is being pursued by several research groups[32−42] but still requires significant additional effort along a number of directions to be fully practical, for example, the design of a continuum solvation model that is compatible with the polarizable force field.[43,44] If a polarizable force field is used to represent the protein, then the internal dielectric clearly should be unity, as all internal polarization effects are now being modeled explicitly.

In the context of a fixed charge force field, possible solutions of the internal dielectric problem must involve heuristic approximations. We describe one such approximation below and then demonstrate that significant improvement in both side chain and loop prediction is obtained with the use of a few adjustable parameters, which assume physically reasonable values. Furthermore, simpler approaches, such as increasing the internal dielectric constant from one to a higher value, lead to results that are qualitatively inferior.

The basic idea is to vary the value of the internal dielectric constant as a function of the interacting atoms. In the GB model, the total electrostatic free energy is expressed as the sum of the Coulomb interaction and the generalized Born solvation term

$$G_{es} = \frac{1}{2}\sum_{i<j}\frac{q_i q_j}{r_{ij}\epsilon_{in}} - \frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{sol}}\right)\sum_{ij}\frac{q_i q_j}{f_{GB}} \qquad (1)$$

where

$$f_{GB} = \sqrt{r_{ij}^{\,2} + \alpha_{ij}^{\,2} e^{-D}}$$

and

$$\alpha_{ij} = \sqrt{\alpha_i \alpha_j}, \quad D = \frac{r_{ij}^{\,2}}{(2\alpha_{ij})^2}$$

The α_i's are generalized Born radii. In our variable dielectric model, the internal dielectric ϵ_{in} depends upon the pair of atoms that are involved in the specified electrostatic interaction. We write this explicitly:

$$G_{es} = \frac{1}{2}\sum_{i<j}\frac{q_i q_j}{r_{ij}\epsilon_{in(ij)}} - \frac{1}{2}\left(\frac{1}{\epsilon_{in(ij)}} - \frac{1}{\epsilon_{sol}}\right)\sum_{ij}\frac{q_i q_j}{f_{GB}} \qquad (2)$$

Note that this newly defined $\epsilon_{in(ij)}$ enters both the solvation terms and the Coulomb interaction terms. We use a residue-based parametrization with the variation confined to side chain atoms of charged residues, i.e., we assign different dielectric constants for charged side chains, while all the backbone atoms and neutral side chains still use the dielectric 1. There are several reasons for not changing dielectric constants of backbone atoms: (1) We want to treat the backbones consistently with both charged and neutral residues, as they have the same parameters (charge, Lennard-Jones, etc.) independent of the residue type in the OPLS-AA force field. (2) Although the individual atoms in the backbone carbonyl and the amine group are significantly charged, the backbone as a whole is neutral, hence the argument that an appropriate polarization has already been incorporated via optimization of the charges in liquid-state simulations applies. (3) Experiments with structure predic-

Table 1. Internal Dielectric Constants Used in a Variable Dielectric Model[a]

residue	Lys	Glu	Asp	Arg	His	other
dielectric	4	3	2	2	2	1

[a] The new values other than 1 are only assigned to the side chain atoms. For a specific interacting pair, the internal dielectric uses the larger one of the two values associated with the two atoms.

tions prove it is a better choice to use unity dielectric for backbone atoms than varying it with residue type (data not shown).

For the inter-residue pair interaction, in the present paper we employ a simple rule in which the higher of the two "residues-based" dielectric constants are used. If both interacting atoms are neutral, then according to the arguments described above, the internal dielectric constant is set to 1. On the other hand, if one or both of the atoms belong to the side chain of a charged residue, then the higher internal dielectric associated with the two atoms is employed. The adjustment of the residue-based dielectrics is accomplished through the improvement of the single side chain predictions. We first use a uniform dielectric of 1 to 6, 8, and 20 for the entire protein and determine which value yields the most accurate predictions (measured by the percentage accuracy, see Results) for each residue and choose this value to be each residue's dielectric. We did not try a finer parametrization because the structure prediction is not very sensitive to the slight changes in the dielectric constant any smaller than 1. Then, in the variable dielectric model where the combining rule is applied for inter-residue interactions, we further adjust these dielectric values (by a maximum of ± 1) to maximize the overall accuracy of the single side chain predictions of all 11 polar or charged residues. The optimized set of values is presented in Table 1; the results obtained using these values are given in section V. Considering there are 2178 single side chain prediction test cases and only 5 adjustable parameters, it seems unlikely that the results are due to gross overfitting. The significant improvement of loop predictions by the variable dielectric model (see section V) also provides an independent validation test. A more sophisticated optimization could be employed and might yield better results; some possibilities along these lines are considered in the Discussion section.

Before examining detailed numerical results, it is useful to obtain an intuitive physical feeling for the results of the variable dielectric model outlined above and to see whether such results will move side chain predictions in the qualitatively correct direction. For charged residues a well-known, fundamental problem of dielectric continuum models is a tendency to form salt bridges considerably more frequently than is observed experimentally; this was demonstrated, and discussed, in our previous work.[45] In contrast, hydrogen bonds between neutral residues do not display an extreme bias one way or another. Thus, the hope would be that the variable dielectric model can reduce the frequency of salt bridges, while having minimal impact on neutral−neutral hydrogen bonding.

A simple physical argument suggests that this will indeed be the case. Take the case of the interaction of a lysine residue with the surrounding atoms of the protein. The internal dielectric refers to all of the atoms surrounding the lysine. The polarization of these atoms creates a reaction field around the charged atoms of the lysine group—not as large as the reaction field from water but larger than would be observed if the internal dielectric constant were unity. This reaction field then has an unfavorable interaction with the hydrogen-bonding partner of the lysine, e.g., a carboxylate group. Similarly, the carboxylate group has a reaction field around it that has an unfavorable interaction with the lysine. It is these reaction fields that reduce the magnitude of the effective interactions between the two groups, as in the case of charged groups in water.

This effect is most easily seen from eq 1 in the limit where the Born alpha values of the interacting atoms become very large, as would be the case for a significantly buried salt bridge. In the large alpha limit, the second term in eq 1 becomes negligible, and one is left with only the first term for the interaction energy; this term divides the Coulomb interaction by the internal dielectric constant. In this way, salt bridges become properly energetically disfavored using the variable dielectric model, due to the increased reaction field of the protein surrounding each component of the salt bridge. This diminishes the number of unphysical salt bridges as is demonstrated in more detail below.

For other, more complicated cases, in which alpha is not assumed to be much larger than the separation distance between the interacting groups, the analysis of increasing the internal dielectric becomes more complicated, but the basic physics (creation of a reaction field due to the protein in response to the electric fields from the interacting groups) is unchanged, and the adequacy of the quantitative treatment of this effect by our simple, variable dielectric approximation must be judged by the quality of the results for loop and side chain predictions as presented below.

Now consider the case of two serines interacting with each other, for which, in computing the electrostatic free energy via eq 1, we use an internal dielectric of unity. There is no enhancement of the reaction field surrounding the −OH groups, because the remainder of the protein force field was derived based on a neutral, hydrogen-bonding environment. In contrast, the charged groups will produce a reaction field in the protein in excess of what is incorporated into the force field, because the field exerted on the neighboring protein atoms is in excess of what was used in the parametrization of the model. The empirically tuned variable dielectrics represent a crude, but apparently quite useful approximation to the magnitude of this excess. The variation in the reaction field differential with Born alpha and other geometrical parameters of a given structure is implicit in eq 1; this apparently corresponds well enough to physical reality that substantial improvements in both side chain and loop predictions are obtained. It is worth noting that alternative empirical "fixes" such as changing the dielectric radius of various charged atoms did not yield significant improvements in structural predictions in our experiments (data not shown).

Implementation of the variable, residue-based dielectric model as described above is relatively trivial; the constant internal dielectric used previously is replaced by a variable determined by looking up the appropriate value in Table 1.

The extensive tests, carried out and described in section V, are considerably more time-consuming but necessary to evaluate the performance of the model, which, given its approximate and heuristic character, cannot be rigorously inferred from the theoretical arguments made in this section.

B. Increasing the Speed of PLOP. A number of software optimizations have been included in the current version of PLOP after extensive profiling of the previous versions. Some improvements involve simple steps to avoid unnecessarily expensive copying of large arrays and string compares. More substantially, all function calls and most conditionals have been removed from inner loops of the gradient and energy code. Instead of conditionals for the various solvation types and corrections being placed in the inner loops, separate inner loops for different solvation conditions are generated automatically with pseudocode. Any atom pairs requiring special additional correction terms (such as the hydrophobic pair term introduced previously[4]) are now placed into separate neighbor lists, removing the need for conditionals within the inner loops.

After the inner loops were optimized, the generation of the SGB surface, the integration over this surface, and the determination of Born alphas consumed the most time in both minimizations and loop predictions. A number of improvements to the surface generation and integration code have been performed to eliminate unnecessary checks and duplicated calculations that occurred when only some parts of the surface changed. Additionally, in the intermediate steps of side chain optimization, Born alphas are only updated within twice the solvent radius (1.4 Å for implicit water) of any moving atoms. This reduces the time for calculations done for intermediate steps where only approximate energy evaluations are necessary.

A previously implemented correction to the Generalized Born energy due to Ghosh et al.[8] with the aim of improving the consistency between GB and PB results for protein structures was determined to almost double the time required to determine the surface integral, taking 20−30% of the total time of a loop prediction. A set of side chain and loop prediction tests determined that this correction only negligibly affected the prediction results (average RMSDs differ less than 0.05 Å) and has therefore been removed from the code.

One new algorithmic improvement in the current implementation of PLOP is the replacement of residue-based cutoffs with dipole based cutoffs. In the previous versions of PLOP, potential energy cutoffs were residue based. If any two atoms of a residue were less distant than the specified cutoff, all atoms on those two residues were treated as interacting. Neutral−neutral residues had a first cutoff, neutral-charged residue interactions had a second cutoff, and charged−charged residue interactions had a third cutoff. This leads to an undesirable situation where extremely small changes in structure can result in a significantly larger change in energy for short cutoffs as residues move in and out of the cutoff distance. This residue-based approach can therefore cause instabilities with the multiscale minimization algorithm used in PLOP. In this scheme, inner loops of the minimization use only a short-range cutoff, and the long-range gradient is approximated as a constant. When the neighbor list is updated, however, entire residues may have moved into or out of the cutoff, changing the direction of the minimization significantly. Relatively long short-range cutoffs were therefore required; with cutoffs less than 8 Å, minimizations would frequently become numerically unstable. A default short-range cutoff of 10 Å cutoffs for neutral−neutral residues and neutral-charge residues and 15 Å cutoffs for charge−charge residues was determined to be safe for adequate convergence.

In the PLOP implementation presented in this paper, distance-based interaction cutoffs are still present but are dipole based, instead of residue based. Atomic charges are decomposed into formal charges and dipoles. As an example, we examine the case of a hypothetical neutral methane. If we assign hydrogens a partial charge of 0.1, the central carbon must have a partial charge of −0.4 (all charges are for illustrative purposes and not meant to reflect the actual force field used for predictions). We now represent the partial charges, instead of being atom based, being bond based, with each bond having positive and negative charges equal in magnitude at each end. This methane molecule then consists of four C−H dipoles, each of magnitude 0.1, with the negative poles toward the carbon and positive poles toward the hydrogen. For a hypothetical ammonium ion, with a total charge of 1.0, partial charges of 0.2 on the hydrogens, and a partial charge of 0.2 on the nitrogen, it would be represented as a formal charge of magnitude 1.0 centered on the nitrogen and four dipoles of magnitude 0.2, with negative poles toward the nitrogen. For a neutral molecule, there exists a unique decomposition with the exception of ring systems; a choice is made in the ring systems to minimize the magnitudes of the resulting dipoles.

The atoms of two dipoles interact only if all four atoms are within the cutoff distance of each other. For example, imagine two such methane molecules interacting. Suppose that all pairs of atoms between the two methanes are within the cutoff, with the exception of one hydrogen on methane A and one hydrogen on molecule B whose distance is slightly greater than the cutoff. These two hydrogens do not interact, and because atoms of these two dipoles do not interact, neither do the other atoms in the dipole. The partial charge of both carbons is now the sum of only the three dipoles, so it becomes 0.3. This partial charge is only with respect to the pair, meaning that the effective pairwise partial charge must be determined between all pairs of atoms. If there are two hydrogens on molecule A that are further than the cutoff from one hydrogen on molecule B, then the carbon on methane A would have partial charge 0.2, and the carbon on molecule B would have partial charge 0.3. This method of determining the electrostatics cutoffs has a useful property that the total sum of the product of all charge pairs in the system $q_i q_j$ is independent of the cutoff, at least for those charges that are the result of sums of dipoles, not formal charges. This results in significantly smoother changes in the energy with respect to changes in structure than an abrupt atomic cutoff or even a residue-based cutoff. The basic algorithm, as described up to this point, was implemented in the Macromodel modeling package but has not previously

been published. The additional modifications described below, however, are novel.

Determining the dipole interactions as a function of the distance between pairwise atoms can be somewhat time-consuming and must be redone every time the structure changes. To significantly improve the performance of the method, we group sets of dipoles together. If an atom has only one neighbor, then we call it a leaf. When it has more than one neighbor (or, for practical purposes, it has a formal charge), we call it a trunk. When two trunks interact, then all of their leaves also interact. Therefore, the product of the partial charges for the interaction of any two leaves is zero if their trunks are not within the cutoff. This is equivalent to treating the length of the trunk/leaf dipole as zero for the purpose of determining which dipoles interact (though *not* for the purposes of calculating the energy itself). For example, if we consider two methanol molecules under this system, the C−H dipoles will always interact if the C−C distance is sufficiently close, as will the O−H dipoles, if the O−O distance is sufficiently close. However, the intramolecular C−O dipole will only contribute to the total partial charges involved in the C−C interaction if all four carbons/oxygens intermolecular distances are less than the cutoff. In many cases, the interactions of this form of dipole based cutoffs may die off more quickly as a function of distance that in the original version, as multiple dipoles contributing to a single "trunk" will tend to orient in opposite directions, leading to a smaller average dipole.

In place of neutral−neutral, neutral−charged, and charged−charged residue cutoffs, all dipole−dipole interactions and all Lennard-Jones interactions are truncated with a single dipole−dipole cutoff, with dipole−formal charge and formal charge−formal charge cutoffs treated separately with longer cutoffs. Previously, Lennard-Jones interactions between charged−charged residues or charged−neutral residues as well as the electrostatics of nonpolar moieties were calculated at a larger distance than Lennard-Jones interactions in neutral−neutral sides. Although these deviations from a uniform treatment of the Lennard-Jones terms are relatively small, they can cause inconsistencies when, for example, the protonation state of a residue is changed.[22]

For multiscale minimization, PLOP uses two separate neighbor lists: a long-range list that is used to calculate the full gradient and energy and a short-range list that is used in calculating the short-range gradient and Hessian for the inner loops and preconditioner of the minimizer. The dipole-based cutoff scheme is applied to both sets of cutoffs.

With the new dipole-based cutoff system, the number of atoms included in the short-range energy and gradient evaluation can be reduced significantly, with a relatively small computational cost for the additional bookkeeping. Well converged results can be obtained with cutoffs as short as 6 Å, and the use of even shorter cutoffs is only ruled out because of bookkeeping difficulties with the 1−4 interactions, which must be treated separately. As the short-range cutoffs get smaller, however, the approximation used in the multiscale minimization of a long-range contribution to the gradient which is constant constant during the minimization inner loop becomes less accurate, increasing the number of iterations required to converge minimizations. Short-range cutoffs of 8 Å dipole−dipole interactions, 10 Å for dipole−formal charge interactions, and 16 Å for formal charge−formal charge interactions provide near optimal speed in most instances. Long-range cutoffs of 15 Å, 20 Å, and 30 Å were used. These long-range cutoffs are the same as used for the previous residue-based cutoffs and thus involve somewhat fewer atoms but yield similar relative energies between structures with either cutoff scheme.

IV. Test Set

For the single side chain prediction test, we use the test suite of 30 protein structures from the previously publication.[22] These proteins are selected such that all of them have resolution <2 Å and do not have any serious heavy atom steric clashes or nonpeptide ligands. The pairwise sequence identity is less than 30%. In this work, we add a new screening criterion based on the B-factor to eliminate the noise in the results due to the experimental uncertainty. If any side chain atom has a B-factor greater than a threshold of 40, then that residue will be excluded from our list. We focus on the predictions of 11 polar (and charged) residues since the hydrophobic residues are generally buried and trivial to predict and thus less affected by the solvent model. This yields a total of 2178 single residue side chain structure repredictions.

For the loop prediction targets, we use the combined filtered list in ref 3 for 6, 8, and 10 residue loops. The 13 residue loops are from ref 4. These loop targets were filtered by pH value, B-factor, steric clash, and other criteria to ensure the selection of high-quality structures. In total, there are 99, 65, 41, and 35 targets for 6, 8, 10, and 13 residue loops. The loop target 9-14 of 1xso from the 6 residue list and the target 606-613 of 1gof from the 8 residue list are removed because of serious steric clashes in these structures.

V. Results

A. Single Side Chain Predictions. Single side chain predictions represent one of the simplest tests that can be applied to evaluate the quality of a protein energy model. We focus on comparisons to crystallographic structural data, as opposed to NMR data, as it is not clear whether NMR data for side chains is precise enough to enable robust comparisons at this point in time. To compare realistically with X-ray crystallographic data from the PDB, the calculations must be carried out in the appropriate crystalline environment; many side chains form nonbonded contacts (e.g., salt bridges) with neighboring protein molecules in the crystal. In the calculation, all atoms other than those of the side chain in question are held fixed, the conformational phase space of the side chain is sampled as thoroughly as possible, and the energy model (molecular mechanics potential energy plus free energy due to the continuum solvation model plus some entropic term) is used to select the final prediction. Because the number of degrees of freedom in a single side chain is small as compared to a long loop, it is generally possible to converge the side chain sampling to the global free energy minimum, although in a small number of cases in the data sets, the presence of a

Table 2. Single Side Chain Prediction Accuracy of 11 Polar Residues on a Data Set of 30 Proteins and 2178 Side Chain Targets[a]

residue type	total no.	uniform dielectric 1 (%)	uniform dielectric 2 (%)	uniform dielectric 4 (%)	variable dielectric (%)	variable dielectric with ICDA assignment (%)
ASN	237	71.7	72.6	70.5	75.5	85.7
GLN	161	65.8	65.8	58.4	65.2	85.7
HIS	132	54.5	60.6	58.3	59.8	86.4
ASP	254	83.9	86.2	83.9	87.8	91.7
GLU	193	67.4	74.1	75.6	73.6	79.3
SER	297	77.4	63.0	42.8	80.1	79.1
THR	302	90.1	90.7	88.1	91.7	92.4
LYS	198	62.6	70.7	77.8	76.8	76.8
ARG	171	78.4	77.2	70.8	74.9	77.8
CYS	49	93.9	89.8	89.8	93.9	93.9
TYR	184	88.0	89.7	92.4	91.3	89.7
SUM	2178	76.2	76.3	72.5	79.8	85.0

[a] The accurate prediction is defined as having side chain heavy atom RMSD less than 1.5 Å. The variable dielectric model is compared with the uniform dielectric models assuming different dielectric constants. The last column shows the single side chain prediction results with the ICDA assignment structures using variable dielectric model.

positive energy gap between the predicted structure and minimized native structures indicate that convergence is not, in fact, achieved using our current sampling algorithms.

While single side chain prediction tests are relatively straightforward to execute, it is far from trivial to attain robust prediction of experimental side chain geometries, particularly as the side chain becomes more solvent exposed. When the side chain is buried in the interior of the protein, geometrical constraints leave few alternatives with regard to configuring the side chain in a manner that is compatible with the remainder of the protein, which is kept rigid in the prediction. However, as the degree of solvent exposure increases, the number of plausible alternative conformations also increases. For example, a specific solvent exposed lysine can often form either a salt bridge or a charge-neutral hydrogen bond or remain free in solution, interacting closely with no other protein atoms. In many cases these configurations may be close in free energy and hence difficult for any energy model to discriminate; in other cases, the energy model may make large, systematic errors, incorrectly preferring one type of structure over another. Such solvent exposed cases provide a significant challenge to energy models, one that enables reliable assessment of the accuracy of the model for a wide variety of interactions.

The single side chain prediction results with the variable dielectric model are shown in Table 2 as well as a comparison with a variety of other possible dielectric models. We test our energy model on single side chain prediction of 11 polar and charged residues. The nonpolar residues have relatively small partial charges, and most of them are buried in the interior of the protein. The van der Waals interactions, instead of electrostatic interactions, are often the dominating forces for their conformations. Thus, nonpolar residues are less affected by the solvent model and hence are ignored in this study. We use the root-mean-square-deviation (RMSD)

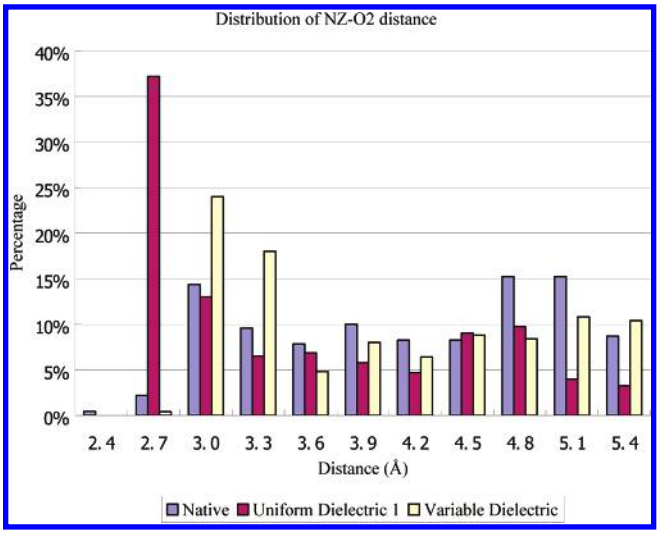

Figure 1. The distribution of distances between the lysine NZ atom and the carboxylic acid O2 atoms. The predictions of uniform dielectric 1 and the variable dielectric model are compared with native structures. The variable dielectric model eliminates the overprediction of salt bridges in the uniform dielectric model.

of all heavy side chain atoms as the accuracy measure (excluding the C^{β} atom which is largely fixed by backbone position). This measure accounts for the positions of the entire side chain and is more suitable for high-resolution comparison than the χ_1 and χ_{1+2} angles. 1.5 Å is chosen as the threshold for an accurate prediction. As Table 2 shows, compared with the default model (uniform dielectric of 1), the variable dielectric model improves all polar side chain accuracies, to varying degrees, except for arginine. The largest improvements come from lysine and two carboxylic acids predictions. The percentage of accurate predictions for lysine increases from 62.6% to 76.8%; this reduces the number of errors in lysine prediction from 74 to 46 or by a factor of 38%. For glutamate and aspartate, the accuracies increase from 67.4% and 83.9% to 73.6% and 87.8%, respectively. This is equivalent to the error reduction of 19% for glutamate and 24% for aspartate. The overall accuracy for these 2178 residues increases by a substantial amount, from 76.2% to 79.8%.

The single largest error eliminated by the variable dielectric model is the overstabilization of salt bridges. This overstabilization problem occurs on many other GB-type models and has been observed in various simulations.[46−49] In our single side chain predictions, a recurring scenario was that the solvent exposed lysines were often predicted to form a salt bridge instead of being free in solution as in the native structure. This clearly occurs because forming the salt bridges receives excessive stabilization energy in the energy model, as compared with being solvated in solution. The new variable dielectric model solves this problem. In Figure 1, we plot the distance distribution of lysine NZ atom and carboxylic O2 atom (glutamate and aspartate) in native structure and predicted structures. The NZ-O2 distances in native structures are relatively flat and show two maxima at 3.0 Å and 4.9 Å, which approximately correspond to the contact minimum and solvent-separated minimum in the

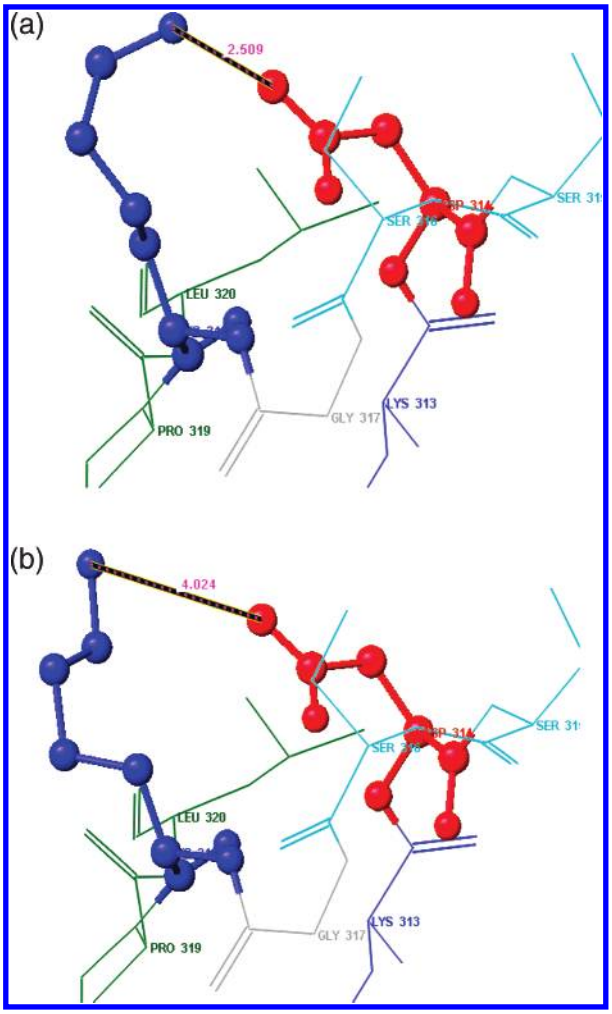

Figure 2. An example of using a variable dielectric model to improve the single side chain prediction. The single side chain prediction on 1ixh Lys318 with a uniform dielectric yields a structure with an erroneous salt bridge between Lys318 and Asp314. The RMSD is 2.82 Å (a). In the variable dielectric model, the lysine is correctly predicted with a RMSD of 0.74 Å (b).

potential of mean force (PMF) between NZ and O2 atoms. The predicted structures by the previously uniform dielectric energy model show a very high population at the distance around 2.8 Å. This overwhelmingly strong attraction between the NZ-O2 atoms leads to a collapse, which is held at a distance of 2.8 Å by the repulsion in the van der Waals term. The variable dielectric model prevents this collapse and greatly diminishes the population of salt bridges. Although not perfect, the NZ-O2 distribution shows two similar maxima as in the experimental structural data.

One specific example of the improvement the variable dielectric model produces is given Figure 2. The single side chain prediction on 1ixh Lys318 with uniform dielectric yields a bad structure with RMSD 2.82 Å. Lys318 and Asp314 form an erroneous salt bridge, and the distance between lysine NZ atom and carboxylic O2 atom is 2.51 Å. Using the variable dielectric model, the lysine is correctly predicted with a RMSD of 0.74 Å. The lysine ammonium group extends into the solvent, and the distance between lysine NZ atom and carboxylic O2 atom is 4.02 Å. Such a NZ-O2 distance is very commonly seen in the crystal structure and is favored because it allows for bridging waters between the two oppositely charged groups, maximizing the hydrogen bonding.

In addition to calculations employing a fixed dielectric of 1 and our new variable dielectric model, we also present results for fixed internal dielectric constants of 2 and 4 in Table 2. Values higher than 4 lead to significantly worse results and are not shown here. These results demonstrate that, as argued above, the use of any single alternative to unity as an internal dielectric does not improve the overall performance of side chain prediction. In particular, it is interesting to note that polar, uncharged side chains such as serine experience substantial degradation in performance as the single dielectric is increased. This is consistent with the hypothesis discussed above that for neutral−neutral hydrogen bond interactions the force field already has appropriate polarization included as a result of fitting to pure liquid-state simulations, and hence the use of a larger internal dielectric for these interactions is in effect double counting.

A complicating factor in using single side chain prediction to evaluate energy models is the dependence of prediction accuracy upon correct assignment of protonation states of the various side chains. For example, if a protonated histidine forms a salt bridge with a carboxylic acid in the native structure, a prediction performed with an unprotonated form of histidine may well prefer an alternative structure. To address this problem, we have developed a protonation state assignment methodology (referred to as Independent Cluster Decomposition Algorithm (ICDA), described in detail in ref 22), which already has been shown to provide substantial improvements in protonation state prediction given a crystal structure as a starting point. The ICDA infers the location of hydrogens in a high-resolution crystal structure based on the heavy atom positions obtained experimentally; it does not imply a complete search of conformational space, as the heavy atom positions are kept fixed during ICDA calculations. Hence, it is unsurprising that when hydrogens are assigned via the ICDA protocol, the native side chain conformer will in many cases be stabilized as compared to incorrect alternatives. However, the ICDA in and of itself is insufficient to produce perfectly accurate single side chain predictions; an accurate energy model is also essential; we illustrate this point by comparing side chain prediction results using the ICDA for both our old and new dielectric models.

In evaluating our new variable dielectric methodology, we perform comparisons with and without first assigning the protonation state by ICDA; as shown in Table 2, the combination of protonation state assignment and improved dielectric model yield substantially better results than either approach used by itself. For the 11 polar residues, the accuracy from the combination is 85.0%. The histidine accuracy increases from 59.8% to 86.4%. In this process, we simply take the structures generated by the ICDA assignment using the fixed dielectric of one and run the single side chain prediction with the variable dielectric model. It is possible to apply the new variable dielectric model into the ICDA algorithm; however, this would be involved with

Table 3. Classification of Prediction Errors for Four Charged Residues[a]

	sampling error (%)	energy error (%)	solution error (%)	hydrogen bond error (%)
Lys	10.9	90.1	43.9	56.1
Arg	23.3	76.7	12.1	87.9
Asp	19.4	80.6	4.0	96.0
Glu	35.3	64.7	21.2	79.8

[a] The results are based on the predictions with variable dielectric and ICDA assignment. The prediction error is defined as having a side chain heavy atom RMSD greater than 1.5 Å. If the energy of the predicted structure is higher than the directly minimized structure, then it is a sampling error, otherwise it is an energy error. All energy errors are further classified into two types. If both the native structure and the predicted structure do not form any hydrogen bond with other residues, then it is defined as a "solution error", otherwise it is defined as a "hydrogen bond error". The distance cutoff for a hydrogen bond is 3.1 Å between the acceptor and the donator, and no angle consideration is involved.

extensive reparametrization of ICDA algorithm, which we decided not to pursue at this point.

Having obtained such a significant improvement in the single side chain predictions, there is still a noticeable percentage of errors, especially for four charged residues: lysine, arginine, glutamate, and aspartate. We classify the errors into either energy or sampling errors, as shown in Table 3. If the energy of the final predicted structure is higher than the directly minimized native structure, then it indicates the sampling is not sufficient. If the energy of the final predicted structure is instead lower than the energy of the minimized native structure, then the error is attributed to the deficiency of energy model. Furthermore, we classify energy errors into two types: solution error and hydrogen bond error. If the side chain in both the experimental structure and the predicted structure does not form any hydrogen bond with the rest of the protein body, we call this type of misprediction a solution error. Otherwise, if the side chain is forming hydrogen bond(s) with the protein body in either the experimental structure or the predicted structure, then the error is designated as a hydrogen bond error. The hydrogen bond error represents a class of error that is more likely to be fixed by further improving the energy model because they are relatively easy to characterize. Table 3 shows that glutamate has the highest sampling error percentage of 35.3%, while arginine is second with 23.3%. This means that although the present sampling algorithm could produce accurate predictions for a majority of the cases, it stills needs to be improved for certain residues. Among the energy errors, the lysine has a very large percentage of solution errors in this study, at 43.9%. This of course is due to the fact that the lysine tends to be fully solvated in the solution. In contrast, most of energy errors of the aspartate are hydrogen bond errors. The characterization and correction of this type of error should have a high priority in order to further improve our energy model.

B. Loop Prediction. We apply the new variable dielectric model on a set of loop predictions ranging from 6, 8, 10, and 13 residue of loop length. We use the two-stage sampling in Jacobson et al.[2] paper for 6, 8, and 10 residue loops and a more powerful yet expensive multistage sampling algorithm for 13 residue loops.[4] The greatly improved accuracies are obtained on all length scales as Table 4 shows (the detailed results are in the Supporting Information) when we compare the variable dielectric model and the uniform dielectric model. The average loop backbone RMSDs (superimposing the rest of the protein) for 6, 8, 10, and 13 residue loops decrease from 0.48 Å, 0.84 Å, 1.27 Å and 2.73 Å to 0.40 Å, 0.79 Å, 0.73 Å, and 1.62 Å, respectively. In ref 4, we introduced a hydrophobic term into the SGB/NP model, which greatly improved the accuracy of long loop predictions. We attributed the success to the correction of absent hydrophobic interaction in the SGB/NP model, which is more prominent in the long loop prediction. Given the substantial advantage of a hydrophobic term on the SGB/NP model, it is important to verify whether it is compatible with the variable dielectric model. Table 4 shows that using the hydrophobic term improves the accuracy of loop prediction on both the variable dielectric model and the original SGB/NP model. The combination of both the variable dielectric model and the hydrophobic term yields the best accuracies of loop predictions with average backbone RMSDs of 0.41 Å, 0.74 Å, 0.76 Å, and 1.08 Å for 6, 8, 10, and 13 residue loops, respectively.

We define the energy gap (EGAP) as the energy of the predicted structure minus the energy of the directly minimized native structure. With the assumption that the native structure well represents the global minimum on the free energy surface, an ideal prediction should yield a reasonably good structure with a zero or slightly negative energy gap. A large negative energy gap with an incorrect structure indicates the energy function is flawed and thus has to be improved. A positive energy gap implies that the sampling is not sufficient; the status of the energy model for a test case of this type is unclear, although in practice such cases usually minimize to the native structure if one can locate it. Since our sampling method is a multiple stage process guided by the energy function, the energy function will bias the sampling to its favorable conformational space. A good energy function could bring the sampling region closer and closer to the native structure and finally find a nativelike structure, while a bad energy function would fail to do that. This difference is more prominent for long loops since sampling is more challenging in these cases. For example, there are a number of sampling errors in the predictions of 13 residue loops using the original SGB/NP model, while other improved energy models (hydrophobic term, variable dielectric, or a combination of both) eliminate the sampling errors, although the same sampling protocol is used (Table 4 shows the average energy gap. See the Supporting Information for detailed information.).

As Table 4 shows, the improvement due to the hydrophobic term when using the variable dielectric model is not as large as the improvement it provided with the original SGB/NP model. For example, when the hydrophobic term is introduced into the SGB/NP model, the RMSD for 13 residue loops decreases from 2.73 Å to 1.29 Å, while the combination of the hydrophobic term with the variable dielectric model reduces the RMSD from 1.62 Å to 1.08 Å. This is because sometimes both the hydrophobic term and the variable dielectric model fix the same problematic cases

Table 4. Average RMSDs and Energy Gaps for the Loop Prediction on 6, 8, 10, and 13 Residue Loops[a]

	uniform dielectric		variable dielectric		uniform dielectric + hydrophobic		variable dielectric + hydrophobic		variable dielectric + optHydrophobic	
	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP	RMSD	EGAP
6 residue	0.48	−4.09	0.40	−2.56	0.46	−4.09	0.41	−3.30	0.39	−3.50
8 residue	0.84	−6.48	0.79	−4.45	0.76	−7.50	0.74	−5.71	0.68	−5.09
10 residue	1.27	−4.96	0.73	−0.77	1.05	−4.38	0.76	−3.29	0.80	−6.23
13 residue	2.73	0.00	1.62	−1.17	1.29	−8.90	1.08	−3.65	1.00	−7.21

[a] The RMSD is the loop backbone RMSD while superimposing the rest of the protein. The energy gap (EGAP) is the energy of the predicted structure minus the energy of the directly minimized native structure. The units for RMSD and energy gaps are Å and kcal/mol, respectively. The first two columns show the results with a uniform dielectric model and a variable dielectric model. The next two columns show the results when these two models are combined with the hydrophobic term. The last column shows the results of our optimization of the hydrophobic term on the variable dielectric model by taking lysines out of the hydrophobic term. Hydrophobic and optHydrophobic represent the original hydrophobic term and the optimized hydrophobic term, respectively.

in the original SGB/NP model. They treat different physical phenomena: the hydrophobic term compensates for the absent hydrophobic interactions and variable dielectric screens for excessively strong charged interactions. However, they can still lead to duplicate effects in terms of generating the delicate balance among various forces that determinate the loop geometry. For example, the reduction of polar (electrostatic) interactions has somewhat similar effects as enhancing the nonpolar (hydrophobic) contributions. Thus sometimes the combination of the hydrophobic term and the variable dielectric does not work as well as either of them works alone. It should be possible to reoptimize the hydrophobic term with the new variable dielectric model to obtain better performance. In a preliminary effort, we exclude the lysine atoms from the hydrophobic energy term, which was originally defined as all heavy atoms with a partial charge less than 0.25 (absolute value) and thus contained some lysine atoms. The results are shown in the last column of Table 4. The average backbone RMSDs for 6, 8, 10, and 13 residue loops are further reduced down to 0.39 Å, 0.68 Å, 0.80 Å, and 1.00 Å, respectively. However, extensive reparametrization would require significant effort and is beyond the scope of this study.

C. Speed Comparison with the Previous Version. We compare the computational efficiency between the latest PLOP version and the version used in the previous publication[4,23] on a variety of tests. The first set of tests is the minimization of 35 13-residue loops. The minimizations start from the native structures and are converged until the norm of the gradient is below 0.001 kcal/mol/Å. We perform the minimization using both vacuum and generalized Born solvation conditions. The minimization using solvent conditions involves a self-consistent procedure in which the Born alphas are held fixed during the course of each minimization, then updated prior to the subsequent minimization, and so on until the energy ceases to decrease by more than 1 kcal/mol over one course of minimization. The 35 loop minimizations in vacuum are on average 3.1 times faster with the optimized code; all other variables such as processor and compiler are kept constant. This speedup comes from the removal of conditionals and function calls from the inner loops of energy and gradient evaluations as well as the reduction of short-range cutoffs due to the dipole-based cutoffs. For the minimization in the solvent, we separate the time spent on the minimization itself and the update of the Born alphas, as the latter requires significantly more time.

Table 5. Computational Costs for Loop Predictions[a]

	CPU time (h)			
	6 residue	8 residue	10 residue	13 residue
mean	4.7	14.4	91.0	333.6
median	3.0	11.3	85.4	277.7
min	0.5	3.1	21.2	87.9
max	71.2	77.6	198.9	1126.4

[a] The CPU time refers to the cumulative time counted as if on a single processor.

The speedup factor for the minimization itself and for the update of the Born alphas are 6.5 and 2.6, respectively. The additional speedup factor for the minimization relative to the vacuum mainly comes from the elimination of SGB correction terms.

With the significant acceleration of minimizations and SGB calculations, the speed of loop predictions is also increased substantially. Single loop predictions for the 65 8-residue loops becomes, on average, 4.5 times faster. However, the multistage sampling protocol also involves steps of constrained loop buildup, which limits loop buildup within a certain distance of a given structure. This process often takes a longer time than the unconstrained buildup, because the effective resolution to sample the backbone library has to be reduced gradually to generate enough number of loop candidates that meet the distance constraint. Sometimes the buildup stage takes a significant percentage of the total time expense. The actual speedup of the full loop prediction is therefore smaller than the speedup of a single PLOP run without any constraint on the loop buildup. Table 5 shows the computational cost of our loop predictions. The average time cost of a 13-residue loop is 13.9 CPU days. Compared with the results in the previous paper,[4] where the average time for 13 residue loops is 31.4 CPU days, the speedup factor is 2.3. Since the multistage loop prediction protocol is highly parallel, the prediction of a 13-residue loop on a midsize cluster of around 32 nodes usually takes 1−2 days.

VI. Discussion and Conclusion

The results in the last column of Table 2 (variable dielectric plus ICDA model) represents a very substantial improvement in the accuracy of single side chain prediction as compared to our previous results, in the third column of Table 2. Comparison with the work of others is difficult because most

papers do not report single side chain prediction results, and because those few that do typically do not incorporate the crystal environment into their predictions, making a fair comparison of the energy models problematic. Table 3 shows that the performance of the energy model is in fact substantially better than what is implied in the most conservative interpretation of the Table 2 data. That is, a nontrivial fraction of the errors reported in Table 2 are due to either sampling errors (which presumably could be fixed by the application of greater computational effort) or (primarily in the case of lysine) the inability to discriminate alternative structures in solution, a task that probably requires a considerably more accurate energy function than discriminating between hydrogen-bonded and non-hydrogen-bonded structures. It is quite possible that these alternative solution structures are very close in energy and hence both well populated in the native state of the protein; furthermore, for many practical applications, it may not matter very much which solution structures are used in the model.

Examination of the remaining hydrogen bond errors suggests further directions for improvement. Preliminary analysis indicates that residual errors in the molecular mechanics force field (such as torsional parameters) make significant contributions to the errors. Improving the force field will require performing suitable high level quantum chemical calculations, in conjunction with the side chain calculations presented here; work along these lines is currently in progress. Overall, it appears as though the goal of developing a robust, accurate side chain prediction method is within reach in the next few years.

The loop prediction results represent a nontrivial test of the side chain optimization effort, with no further adjustable parameters involved. The significant improvements observed in Table 4 validate the methodology independently and suggest that it is useful to adopt the protocol of revising the solvation model (and force field, if necessary) to fit the single side chain prediction data, driven by physical chemistry based models and calculations. The loop prediction results, to our knowledge, represent the best results reported in the literature to date. In combination, the loop and side chain prediction accuracy and robustness should now be sufficient for high-resolution tasks such as structural refinement of homology models, with the goal of enabling structure based drug design starting from the resulting refined active sites. There still remains a very significant challenge to put in place a global sampling algorithm that will enable progress in homology refinement, even assuming that the energy function is of the quality hypothesized above. In particular, loop and side chain prediction involve relatively localized structural changes, whereas homology models typically have errors distributed throughout the structure and hence a purely localized search strategy may not work. The speed improvements in the sampling algorithm reported here would be helpful in adapting our conformational search strategy to address this and other more realistic, delocalized problems.

The success of the variable dielectric model in improving side chain prediction, while useful in and of itself, particularly because of the simplicity of implementation, also suggests that an accurate treatment of polarization is important in achieving quantitative results in biomolecular modeling. A great deal of success has been obtained with fixed charge force fields which incorporate an average polarization as discussed above, but as one imposes more demanding criteria upon the model in terms of accuracy and robustness at the level of microscopic detail, it is unsurprising that this relatively crude approximation is increasingly problematic. The variable dielectric model is itself a crude approximation but does represent an improvement over taking no account of the difference in environment represented by charged (as opposed to neutral) polar groups. A model explicitly incorporating polarization presumably would be better still, but, as mentioned above, this is significantly work to develop and test. Nevertheless, the results presented here should provide strong motivation for efforts in this direction.

Acknowledgment. This work was supported in part by a grant to R.A.F. from the NIH (GM-52018). M.R.S. was supported by an NIH Ruth L. Kirschstein NRSA fellowship. The authors thank Barry Honig and Matt Jacobson for numerous useful discussions.

Supporting Information Available: Detailed loop prediction results. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Jacobson, M. P.; Friesner, R. A.; Xiang, Z.; Honig, B. *J. Mol. Biol.* **2002**, *320*, 597−608.

(2) Jacobson, M. P.; Pincus, D. L.; Rapp, C. S.; Day, T. J. F.; Honig, B.; Shaw, D. E.; Friesner, R. A. *Proteins* **2004**, *55*, 351−367.

(3) Jacobson, M. P.; Kaminski, G. A.; Friesner, R. A.; Rapp, C. S. *J. Phys. Chem. B* **2002**, *106*, 11673−11680.

(4) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins* **2006**, *65*, 438−452.

(5) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc*. **1996**, *118*, 11225−11236.

(6) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J. *J. Phys. Chem. B* **2001**, *105*, 6474.

(7) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem*. **2001**, *23* (5), 517−529.

(8) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983−10990.

(9) Monnigmann, M.; Floudas, C. A. *Proteins* **2005**, *61*, 748−762.

(10) Rohl, C. A.; Strauss, C. E.; Chivian, D.; Baker, D. *Proteins* **2004**, *55*, 656−677.

(11) DePristo, M. A.; de Bakker, P.; Lovell, S. C.; Blundell, T. L. *Proteins* **2003**, *51*, 41−55.

(12) de Bakker, P.; Depristo, M. A.; Burke, D. F.; Blundell, T. L. *Proteins* **2003**, *51*, 21−40.

(13) Xiang, Z. X.; Soto, C. S.; Honig, B. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 7432−7437.

(14) Fiser, A.; Do, R.; Sali, A. *Protein Sci.* **2000**, *9*, 1753−1773.

(15) Xiang, Z. X.; Honig, B. *J. Mol. Biol.* **2001**, *311*, 421−430.

(16) Xiang, Z.; Steinbach, P. J.; Jacobson, M. P.; Friesner, R. A.; Honig, B. *Proteins* **2007**, *66* (4), 814−823.

(17) Dunbrack, R. L. J.; Karplus, M. *J. Mol. Biol.* **1993**, *230*, 543−574.

(18) Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V. *J. Comput. Chem.* **2004**, *25*, 712−724.

(19) Liang, S.; Grishin, N. V. *Protein Sci.* **2002**, *11*, 322−331.

(20) Desmet, J.; Spriet, J.; Lasters, I. *Proteins* **2002**, *48* (1), 31−43.

(21) Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539−542.

(22) Li, X.; Jacobson, M. P.; Zhu, K.; Zhao, S.; Friesner, R. A. *Proteins* **2007**, *66*, 824−837.

(23) Zhu, K.; Shirts, M. R.; Friesner, R. A.; Jacobson, M. P. *J. Chem. Theory Comput.* **2007**, *3*, 640−648.

(24) Nielsen, J. E.; Vriend, G. *Proteins* **2001**, *43*, 403−412.

(25) Yang, A.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. *Proteins* **1993**, *15*, 252−265.

(26) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35* (24), 7819−7833.

(27) Schutz, C. N.; Warshel, A. *Proteins* **2001**, *44* (4), 400−417.

(28) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373−17387.

(29) Gilson, M. K.; Honig, B. *Biopolymers* **1986**, *25*, 2097−2119.

(30) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118* (35), 8452−8458.

(31) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem. B* **1994**, *98*, 1978−1988.

(32) Grossfield, A.; Ren, P.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125* (50), 15671−15682.

(33) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621−627.

(34) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515−1531.

(35) Patel, S. A.; Brooks, C. L. *Mol. Simul.* **2006**, *32* (3−4), 231−249.

(36) Ponder, J. W.; Case, D. A. *Adv. Prot. Chem.* **2003**, *66*, 27−85.

(37) Ren, P.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933−5947.

(38) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672−679.

(39) Rick, S. W.; Stuart, S. J.; Bader, J. S.; Berne, B. J. *J. Mol. Liq.* **1995**, *65*−*66*, 31.

(40) Rick, S. W.; Stuart, S. J.; Berne, B. J. *J. Chem. Phys.* **1994**, *101* (7), 6141.

(41) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237.

(42) Ren, P.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23* (16), 1497−1506.

(43) Maple, J. R.; Cao, Y. X.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1* (4), 694−715.

(44) Schnieders, M. J.; Baker, N. A.; Ren, P.; Ponder, J. W. *J. Chem. Phys.* **2007**, *126*, 124114.

(45) Yu, Z.; Jacobson, M. P.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **2004**, *108*, 6643−6654.

(46) Zhou, R. *Proteins* **2003**, *53*, 148−161.

(47) Zhou, R.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (20), 12777−12782.

(48) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins* **2004**, *56* (2), 310−321.

(49) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2005**, *2* (1), 115−127.

CT700166F