PROTEINS: Structure, Function, and Bioinformatics 65:438–452 (2006)

Long Loop Prediction Using the Protein Local Optimization Program

Kai Zhu, David L. Pincus, Suwen Zhao, and Richard A. Friesner*
Department of Chemistry, Columbia University, New York, New York

ABSTRACT **We have developed an improved sampling algorithm and energy model for protein loop prediction, the combination of which has yielded the first methodology capable of achieving good results for the prediction of loop backbone conformations of 11 residue length or greater. Applied to our newly constructed test suite of 104 loops ranging from 11 to 13 residues, our method obtains average/median global backbone root-mean-square deviations (RMSDs) to the native structure (superimposing the body of the protein, not the loop itself) of 1.00/0.62 Å for 11 residue loops, 1.15/0.60 Å for 12 residue loops, and 1.25/0.76 Å for 13 residue loops. Sampling errors are virtually eliminated, while energy errors leading to large backbone RMSDs are very infrequent compared to any previously reported efforts, including our own previous study. We attribute this success to both an improved sampling algorithm and, more critically, the inclusion of a hydrophobic term, which appears to approximately fix a major flaw in SGB solvation model that we have been employing. A discussion of these results in the context of the general question of the accuracy of continuum solvation models is presented. Proteins 2006;65:438–452.**

Key words: loop prediction; conformational sampling; continuum solvation model; hydrophobic

INTRODUCTION

Loop prediction has become a canonical problem in assessing methods for high-resolution protein structural modeling. Well-defined test cases can be constructed by starting with a high-resolution structure from the Protein Data Bank (PDB), defining a loop region, and predicting the structure in that region while keep the remaining residues of the protein fixed at their crystallographic coordinates. Realistic applications, such as enumeration of alternative low-energy conformations of the loop (as, for example, are frequently seen in flexible active sites such as kinases), or construction of accurate loop conformations in homology modeling, require reprediction of surrounding side chains (and possibly other degrees of freedom) as well as the loop itself. Thus, success in repredicting native loops in the fixed, crystallographically determined environment, is necessary, but not sufficient, to enable useful practical deployment of the methodology.

In previous work,[1] we have introduced a new approach to loop prediction, in which rigorous hierarchical sampling algorithms are combined with a high-quality molecular mechanics force field and continuum solvation model. These methods have been implemented in the Protein Local Optimization Program (PLOP) and were tested on a suite of ~800 loops ranging in length from 4 to 12 residues. Qualitatively improved accuracy was obtained compared to previous efforts at loop prediction, which principally have employed approximate, knowledge-based potential energy functions, as opposed to a model based on an atomic level description of the physical chemistry.

Although the results in Ref. 1 were encouraging, the performance of the method clearly deteriorated beyond a loop length of ~9 residues. Both sampling errors (i.e., cases where the total energy of the predicted structure was significantly higher than that of the minimized, or side-chain optimized native structure) and energy errors (cases where the total energy of the predicted structure was significantly lower than that of the native structure) increased in frequency compared to shorter loops, and the RMSDs from the native loop of both the sampling and energy errors increased in magnitude. Furthermore, the test suites used for assessing performance on longer loops were inadequate in size.

The problems observed in Ref. 1 for long loop prediction are far from unique to that article. Table I[2–5] presents results taken from work by various groups in predicting loops of length 11 or greater. All these approaches use dihedral angle buildup and candidate selection by a scoring or energy function, but they differ in the algorithm details and energy function compositions. A recent article[6] by Monnigmann et al. provides an overview and brief descriptions for the various alternative methods. It should be noted that the results in Table I are not generated on the same test set; however, they do show some common trend. There is a transition of some sort between 10 and 12 residues, which renders the loop prediction problem quali-

---

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

The first two authors contributed equally to this work.

Grant sponsor: NIH; Grant number: GM 52018 (to R.A.F.).

*Correspondence to: Richard A. Friesner, Department of Chemistry, Columbia University, New York, NY 10027. E-mail: rich@chem.columbia.edu

Received 3 October 2005; Revised 27 January 2006; Accepted 12 March 2006

Published online 22 August 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21040

TABLE I. Long Loop Prediction Results of a Variety of Groups

	Fiser et al.[2]	de Bakker et al.[3]	Xiang et al.[4]	Rohl et al.[5]
8 residue	1.89	2.28	1.45	1.45
11 residue	N/A	4.94	3.52	N/A
12 residue	4.41	4.99	3.42	3.62

The accuracies are measured by global backbone RMSDs in Å

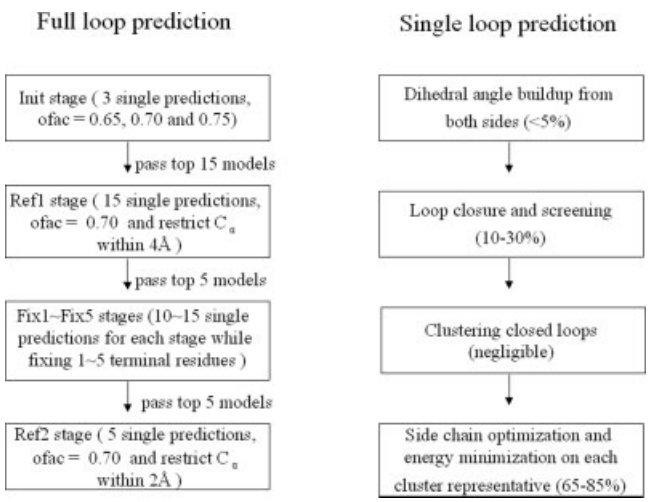

Fig. 1. The flow chart of full loop prediction and single loop prediction. The full loop prediction involves multiple stages, where multiple single loop predictions are executed with different input parameters. The parentheses in single loop prediction indicate the estimated percentage of time expense.

tatively more difficult from the standpoint of both sampling and energetics. This transition is related to the length of the loop compared to the distance between the loop endpoints. When these two quantities are comparable, there are a relatively small number of loop conformations that can connect the end points, and this triaging of phase space via trivial geometrical considerations substantially reduces the degree of difficulty of the loop prediction problem. As the ratio of loop length to end point distance increases, the number of available conformations increases exponentially, and this has two effects: (1) sampling becomes much more challenging, simply due to the rapid growth in the available alternative conformations; (2) the freedom available to the loop can be used to generate false positive conformations exhibiting specific structural motifs favored by the energy model, but not competitive with the native structure if the true free energy of the system were to be evaluated. At a length of 13 residues, it appears as though both of these problems are fully manifested, and the results reported in the literature to date all contain large numbers of qualitatively incorrect predictions, to the point where it seems unlikely that the methodology would be useful in practice in predicting an unknown loop structure of this length. Below 10 residues, Ref. 1 displays few sampling or energy errors, while 11–12 residues constitute a transitional region in which success is possible but not guaranteed.

In the present artcile, we describe a modification of the energy function employed in Ref. 1, along with improved sampling algorithms, which provide qualitatively improved performance in all aspects of long loop prediction. We have constructed a test suite of 104 loops of lengths 11–13, filtering the loops for various criteria as in Ref. 1 (resolution of the protein crystal structure, pH value where the crystal structure was determined, B factor of the loop, interaction with ligands or metal ions, etc.). Sampling errors are virtually eliminated while energy errors leading to large backbone RMSDs are very infrequent, compared to any previously reported efforts, including our own results in Ref. 1. The resulting methodology not only provides a practical approach to accurate prediction of long loops, but also yields insight into the sources of error in continuum solvation models, and the physical interactions controlling loop geometries.

The article is organized as follows. In the Materials and Methods section we provide a brief review of our hierarchical sampling methodology, energy model, and implementation in the PLOP software, and then describe the new energy model and sampling algorithms that form the basis for the present work. In the Results and Discussion section, we first present the results of applying this model and sampling methodology to our long loop prediction test set, and then examine the implications of these results with regard to the accuracy of the potential energy functions and the solvation model. Finally, in the Conclusion, we summarize our results and discuss future directions.

MATERIALS AND METHODS

Sampling Algorithm

Our algorithm for predicting equilibrium loop geometries is a modification of our previous work presented in Jacobson et al.[1] We first provide an overview as to how the new approach is related to our previous algorithm. We then discuss the overall methodology in detail, reviewing the methodology of Ref. 1 as appropriate.

In our present work, a single loop prediction corresponds to the execution of the Protein Local Optimization Program (PLOP) once. Information about PLOP can be obtained from http://francisco.compbio.ucsf.edu/~jacobson/. Our full loop prediction algorithm involves multiple executions of PLOP, enabling additional sampling effort to be focused on loop subsections that are constrained to lie in regions of phase space that have previously been identified as promising. Figure 1 shows the dual scheme of single and full loop predictions. The additional sampling effort allows the loop subsections to be successfully refined to higher accuracy, without requiring an exponential increase in computation time, and without requiring a major rewriting of the core algorithms of PLOP. The multiple execution strategy is readily implemented via PERL scripting, which not only reduces coding effort, but also enables modified strategies to be easily prototyped and experimented with (as, e.g., may be required when loops in the 15–20 residue range are investigated).

Thus, there are two aspects of the sampling algorithm that must be described to characterize the full loop predic-

tion methodology. The first refers to the way that a single loop prediction builds up many candidates, screens and clusters them, and picks out the representatives for scoring and ranking; the second to our procedure that utilizes a variety of ways to sample the low-energy region of conformational space via many PLOP calls in which the options for each execution vary. In what follows, we shall retain the terminology defined above; single-loop prediction will designate a single PLOP execution, while full-loop prediction will designate the ensemble of multiple executions, clustering, filtering, etc., embodied in the PERL script discussed above, yielding the final predicted structure.

Single-Loop Prediction

The method used for single-loop prediction (as defined above) is identical to the loop prediction algorithm described in Ref. 1 (although some parameters of the algorithm are modified from the default values of Ref. 1 in different stages of the full loop prediction methodology, as is discussed below, and the energy model employed in the present article is different). Loop prediction in PLOP is accomplished via an ab initio construction procedure which, at the limit of highest resolution, exhaustively searches the phase space of possible loop geometries connecting the two loop stems. The method 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 that are evaluated via minimization of an all atom molecular mechanics energy function and continuum solvation model.

Each execution of PLOP is composed of four stages: buildup, closure, clustering, and scoring. Exponential scaling is tempered through an adaptive buildup and through screening technology employed in both the buildup and closure stages. Crystal unit cells are explicitly reconstructed by using the dimensions and space group reported in the PDB files. The simulation system consists of one asymmetric unit (which may contain more than one protein chain) and all atoms from other, surrounding symmetric units that are within 30 Å. Every copy of the asymmetric unit is identical at every stage of the calculation; that is, if the conformation of a side chain is modified, all copies of the side chain in the simulation system are updated simultaneously. In the following subsections we briefly review the elements of a single loop prediction, highlighting only those details relevant to the current work. For more details, the reader is referred to Ref. 1.

Buildup

"Buildup" refers to the generation of an initial set of loop conformations that will be passed on to the subsequent three stages. Our buildup is ab initio; it does not rely on the identification of loop fragments from large numbers of solved structures in the Protein Data Bank. The cornerstone of our loop-sampling methodology is dihedral angle search, which is conducted via "rotamer libraries" for backbone dihedral angles (i.e., discretized versions of the well-known Ramachandran plot). To obtain the dihedral angle libraries, we have used a large (>500), nonredundant database of high-resolution (<2 Å) protein crystal structures and recorded every backbone dihedral angle. The dihedral angles were then binned every 5°, and every (ϕ, ψ) combination that appeared more than five times in the database was included in the backbone library. The resultant library, at 5° resolution, contains 747 (ϕ, ψ) combinations for Gly, 215 for Pro, and 866 for all other residue types. The high resolution of the libraries ensures that discretization error does not fundamentally limit the achievable accuracy of a prediction. Unfortunately, it also magnifies the exponential scaling of conformational space and computational expense. We employ an effective sampling resolution to alleviate the problem. Consider the set of (ϕ, ψ) states available to a particular residue, the effective resolution simply guarantees that the distance between any two states in this set is greater than some cutoff. The effective sampling resolution is chosen in an adaptive manner through the following procedure. First, the minimum and maximum number of loops to be generated are specified. The minimum number of loops is 2^n where n is the number of residues in the loop. The maximum number of loops is 10^6. The effective resolution is then decreased from a very coarse value until the number of conformations generated is intermediate between the minimum and maximum parameters. Due to numerous factors (e.g., distance between loop stems, differing environments, amino acid composition of loop, etc.), the number of loop conformations generated for a given target would vary substantially without the use of this procedure. This would, in turn, create difficulties, in terms of memory allocation, algorithmic speed, and effectiveness of the clustering algorithm (described below). Our procedure eliminates this concern.

Screening is primarily accomplished through the use of an "overlap factor" (*ofac*). The overlap factor is defined as the ratio of the distance between two atom centers to the sum of their atomic radii. Multiplying the overlap factor by the sum of the van der Waals radii for a given pair of atoms yields a cutoff. A clash occurs when the distance between the atoms is less than the cutoff. The default ofac of 0.70 is used in the PLOP, but any other values can be prespecified, as we do in the initial stage (see below). Setting the ofac too high can lead to the rejection of loop structures that are close to the native. Setting the definition of steric clash too low, on the other hand, causes a larger number of loop candidates to be generated.

Closure

The buildup procedure continues independently from both sides of the loop up to the Cα atom on the closure residue. Then, all pairs of loop fragments built from the two sides are identified, which have the closure Cα atoms within 0.5 Å of each other. Averaging the positions of the closure Cα atom from the two fragments and adding the Cβ, Hα, and side-chain atoms to the closure residue using standard geometries generates a closed loop. Additional screens are employed at this stage to avoid retaining loops with unacceptable geometries at the site of the closure.

First we check the N–Cα–C angle and backbone dihedral angles on the closure residue. This step is to ensure N–Cα–C angle fall around the ideal value (111.1°) and the dihedral angles within the allowed regions of Ramachandran plot. Then the steric clashes between the two halves of the loop are checked. Because the halves are generated independently, there is no guarantee that they are compatible with each other. Finally, the side chain rotamer library is scanned to ensure at least one conformation of the side chains can fit on the closure residue without steric clash with the body of the protein.

Clustering

Even with the use of our screening technology, the buildup and closure can generate tens of thousands of loops. Optimizing and scoring every candidate with a high-resolution energy model is prohibitively expensive. Furthermore, many of the loops would ultimately optimize into similar structures. Therefore, we use a clustering algorithm to select a representative (and hopefully nonredundant) set of structures for energetic scoring. We employ the K-means algorithm, as implemented by Hartigan and Wong.[7,8] The computational expense of the algorithm scales linearly with the number of items (loops) to cluster. However, unlike clustering algorithms that calculate similarity between every pair of items (and thus scale quadratically), the number of clusters must be specified in advance for K-means algorithm. The number of clusters that we use has been chosen empirically to give satisfactory results for a large number of test cases and scales linearly with loop length. Specifically, the default value, used in all the tests performed here, is four times the number of loop residues. Used in this way, the algorithm rarely requires more than a few seconds of computational time, even with tens of thousands of loop structures. Cluster representatives are chosen according to their distance from the cluster center, with more central representatives chosen first.

Optimization and scoring

Our algorithm relies on an all-atom energy function to discriminate between proposed conformations. Specifically, we rely on the use of an effective potential composed of an OPLS all-atom force field,[9–11] the Surface Generalized Born model of polar solvation,[12] an estimator for the nonpolar component of the solvation free energy developed by Gallichio et al.,[13] and a number of correction terms as detailed in Ghosh et al.[12] and in Jacobson et al.[11] In the present work, we have augmented this energy function by incorporating a new hydrophobic term adapted from the ChemScore[14] scoring function, which has been successfully used to describe the hydrophobic contribution to the binding free energy between ligands and protein receptors. Addition of the new hydrophobic term results in dramatic improvement of prediction accuracy for long loops, as will be shown in detail below. The rationale for the improved performance is examined at length in the Discussion section.

Loop structures chosen for scoring are first subjected to side-chain optimization and complete minimization of the backbone plus side chains of the loop. Our algorithm for side-chain optimization is described in more detail in Jacobson et al.,[15] but is summarized here. Sampling is accomplished by using a highly detailed (10° resolution) rotamer library constructed by Xiang and Honig[16] from a database of 297 proteins. The method we use for the combinatorial optimization is also adapted from the method of Xiang and Honig,[16] which is similar in spirit to earlier work by Bruccoleri and Karplus.[17] In brief, all side-chains are initially built onto the fixed backbone in a random rotamer state, and then each side chain in the protein is optimized one at a time, holding the others fixed. The procedure is iterated until no side chains change rotamer states. After convergence is achieved, the complete loop (side chains plus backbone) is energy minimized in Cartesian coordinates to remove any remaining clashes, and to obtain a reliable estimate of the energy that can be used to fairly compare the energies of the diverse representative cluster members generated in the previous steps, using a rapid, novel multiscale minimization algorithm (Jacobson and Friesner, unpublished results).

Full-Loop Prediction

A full-loop prediction differs from a single-loop prediction in that it consists of a series of multistage, parallel single-loop predictions with varying input parameters. A typical full-loop prediction is composed of eight stages, which are each briefly described in what follows.

Initial stage

The first stage is denoted as the *initial stage* (abbreviated init). The initial stage is composed of three loop predictions executed simultaneously (i.e., three PLOP executions) with three different overlap factors: 0.65, 0.70, and 0.75. The use of multiple ofacs helps to reduce the sensitivity of the results to the definition of steric clash and increase the variety of models generated in this stage. This stage serves to provide a rough guess for the next stages.

After the initial stage, the five lowest energy nonredundant structures are retained from each run; this would give 15 models (e.g., predicted structures). To be identified as nonredundant, none of the structures retained are permitted to have a global backbone RMSD <0.70 Å from any of the other structures retained. The nonredundant parameter is user-adjustable, but is constant throughout the full prediction protocol. The retained structures are passed to the first constrained refinement stage.

The first constrained refinement stage

The first constrained refinement stage (abbreviated. Ref. 1) consists of 15 single loop predictions executed simultaneously. In this stage, each model retained from the previous round is subjected to further sampling using a Cartesian constraint on 4 Å on each Cα atom. The constraint has the effect of allowing us to sample more finely around the low-energy basins identified in the initial stage. Upon the completion of this stage, the five nonredundant best models (e.g., with the lowest energy scores) are

identified from all the models generated from this stage and the initial stage, and passed to the next stage.

The fixed stages

In the "fixed" stages, we predict the structure of short fragments of the target loop while other parts of the target loop are fixed on their given conformations. The motivation for the fixed stage is based on the observation that short-loop predictions have less sampling errors than longer loops. The overall implementation of the fixed stage protocol actually consists of a sequence of stages. Consider a 13-residue loop prediction as an example. First, two 12-residue loop predictions are performed on each model obtained from the previous stage. There will be one residue that needs to be fixed in either side; each case is run via an independent single PLOP execution. This stage is denoted as *Fix1*. After this stage, a few best models will be selected from this stage and all the previous stages and retained for the next stage. Then we do three 11- residue loop predictions on each model while keeping the other two residues fixed. This stage is denoted as *Fix2*. Three predictions for each model are required because there would be three possible positions for an 11- residue continuous fragment. Next, four 10-residue loop predictions will be implemented simultaneously on each model in the next stage. There will be three residues fixed and the stage is denoted as *Fix3*. In principal, this process could go on all the way up to one-residue loop prediction. We find in practice that five such stages are sufficient to get converged results, that is, the energy and RMSD of the best prediction do not change much after that point, at least for loops of length 13 or shorter. For 13-residue loop prediction, the *Fix5* stage corresponds to a set of independent eight-residue single-loop predictions.

There is an important feature in the implementation of these short-fragment predictions, which is different from a normal single loop prediction. In a normal single-loop prediction, the loop candidates are clustered and one representative from each cluster is side chain optimized. This side-chain optimized structure will be ranked together with the representatives from other clusters. However, we do two side-chain optimizations for each cluster representative in our short-fragment predictions: we not only optimize the side chains in the specified short fragment, but also optimize the side chains for the full-loop target. Either structure from these two optimizations with the lower energy will be picked out as the optimized representative and participate into the ranking with the representatives from other cluster. This procedure has the advantage of possibly fixing the mispacking of side chains in the "fixed" region.

During all the successive fixed stages, we always carry out a few parallel runs to maintain the diversity of the sampling pool of loop candidates. However, the number of independent single-loop predictions increases as the fixed stages move on, and this will cause a substantial increase in computational expense if a large number of loop candidates are retained at each stage. For example, the *Fix5* stage requires six independent single executions, and if we keep five parallel models, there will be 30 PLOP executions for this stage. To improve the sampling efficiency, we decrease the size of parallel runs gradually. For *Fix1* stage, these are five parallel runs; for *Fix2* stage, these are four parallel runs, and so on. This strategy reduces by half of the PLOP executions required in the fixed stage and works well for most of our targets. There are a few cases where this algorithm is insufficient, and it is necessary to retain additional candidates further along the optimization pathway. In further work, we intend to refine the methodology so as to preserve robustness while reducing the sampling effort. For the present, this truncated protocol is, in fact, robust for shorter loops; we would recommend the full protocol (i.e., without reducing the size of parallel runs) for the longer loops (12 residues or more) if the goal were to minimize the possibility of an erroneous prediction. In the results reported below, the full protocol is employed as needed for longer loops.

The second constrained refinement stage

This stage is essentially identical to the first constrained refinement stage; only the Cartesian constraint for Cα atoms is changed from 4 to 2 Å. In practice, this stage does not help much to obtain better predictions. This stage is retained because we want to keep some consistency with Ref. 1, where two successive constrained refinement stages are implemented. Our new sampling algorithm inserts multiple fixed stages between the two constrained stages and they prove to be very effective.

New Hydrophobic Term

In the vast majority of continuum solvation models used in protein modeling, the "hydrophobic" or "nonpolar" component of the model is constructed using the accessible surface area of the atoms comprising the system. These terms have typically been fitted to reproduce solvation free energies of small molecules, for example, hydrocarbons, in aqueous solution. PLOP employs a modification of this type of term, developed by Levy and coworkers,[13] which can successfully reproduce experimental results for cyclic, branched, and linear alkanes, in contrast to simple surface area terms which, if fitted to linear alkane data, fail for branched and cyclic molecules.

However, this modification does not address the question of whether the correct description of hydrophobicity in a large, complex structure like a protein might be qualitatively different from that required to understand small molecules, with generally convex surfaces, in bulk solution. In view of the large energy errors reported in Ref. 1 for long loops, we decides to explore this possibility by investigating the use of hydrophobicity models developed in the context of protein–ligand docking. One type of model that has had significant success is based on modeling displacement of waters in hydrophobic regions by the ligand by summing over pairs of lipophilic protein and ligand atoms. The amplitude and distance dependence of the pair term has been optimized by fitting to experimental data on the binding of small molecules to proteins of known structure, for example, complexes available in the

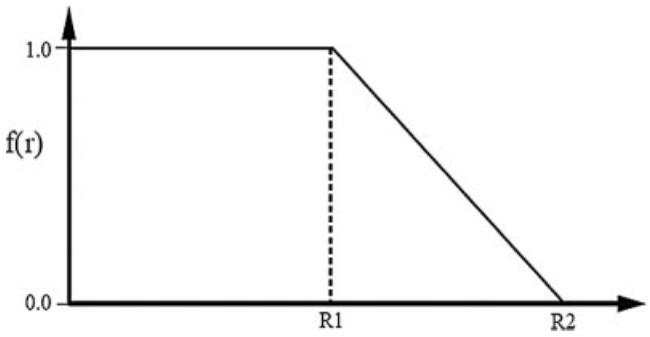

Fig. 2. The function form of hydrophobic term. R1 is the sum of two van der Waals radii of atoms *i* and *j* plus 0.5 Å; R2 is equal to R1 plus 3 Å.

Protein Data Bank (PDB). Such a term is aimed at modeling the free energy that is gained by placing a hydrophobic ligand group (e.g., hydrocarbon chain or phenyl ring) into a hydrophobic pocket of the receptor. The analogy in loop prediction would be placing hydrophobic side chains of the loop into hydrophobic pockets that exist in the protein structure when the loop is removed. This process should be highly favorable in free energy, and captured plausibly (although not perfectly) by the empirical terms developed for protein–ligand docking calculations.

We have chosen to adapt our empirical hydrophobic function from the ChemScore[14] scoring function, which is used to describe the hydrophobic contribution to the binding free energy between ligands and protein receptor in the GOLD docking program;[18] a modified version of ChemScore is also employed in our own Glide[19,20] docking program. We used the linear functional form of ChemScore, reproduced in Equation (1) below:

$$\Delta G^{\text{tot}}_{\text{hydrophobic}} = \Delta G_{\text{pair}} * \sum_{i,j} f(r_{ij}),$$

where *i* and *j* sum over all hydrophobic atoms. A plot of the falloff of $f(r_{ij})$ with distance is presented in Figure 2. However, there are several important modifications in our implementation. Our definition of hydrophobic atoms includes all carbon and sulfur atoms with absolute value of partial charge less than 0.25 unit charge. We also partition the hydrophobic interactions into two parts: the "self" part represents the hydrophobic interactions between all hydrophobic atoms of the target loop region; the "inter" part represents the hydrophobic interactions between the hydrophobic atoms in the target loop and in the rest of the protein (but do not include the atoms from symmetry copies, this point will be discussed later). We choose the prefactor ΔG_{pair} as −0.5 kcal/mol, but for the "self" part this prefactor is scaled by a factor of 0.5. Currently this term, together with the OPLS/SGBNP energy, is used in the screening and scoring of structures, but not used in the minimization where the gradient of energy function is required.

The coefficients specified above have not been fully optimized; rather, a few values have been investigated, and the one that yields highly satisfactory results is used, as will be shown below. However, there are, in fact, a few cases where the RMSD of the loop becomes worse due to addition of the hydrophobic term, suggesting that further optimization of the parameters would be desirable. This, however, requires larger data sets (more cases where the hydrophobic term causes problems) and carrying out numerous loop prediction runs, necessitating the expenditure of large amounts of CPU time. We intend to pursue such optimization in future development efforts; for now, the results reported below clearly represent a qualitative improvement over the original energy model.

Test Set

The results of our previous work suggested that our original loop prediction algorithm was inadequate for loops of greater than 10 residues in length. This conclusion was anecdotal because the 11 and 12 residue filtered subsets used in that study were too small. The small size of the subsets prevented us from extracting meaningful statistics. To remedy this deficiency, we have constructed a new database containing 104 target loops. The target loops are partitioned into three subsets. Each subset corresponds to a different size loop; the first subset is comprised of 38 loops of 11 residues in length, the second is comprised of 31 loops of 12 residues in length, and the third is comprised of 35 loops of 13 residues in length.

Selection Criteria

Proteins from which to select loop targets were found using the PISCES[21] Web server (http://dunbrack.fccc.edu/pisces.php). The following criteria are used to ensure the selection of high-quality structures.

1. The sequence identity between any two proteins has to be ≤40%.
2. The resolution of a structure has to be <2.0 Å.
3. Only crystal structures are selected.
4. Structures with only Cα coordinates are excluded.
5. The maximum allowable *R*-factor is set to 0.25.

Target loops are subsequently chosen from the structures found by the server. The criteria used to select target loops are very similar to the criteria used to filter target loops in our previous work. We outline these criteria below:

1. The pH of the crystal has to fall in the range of 6.5 to 7.5.
2. The average temperature factor of atoms within the loop has to be <35.
3. The minimum overlap factor for any loop-atom has to be ≥0.70.
4. A selected target loop must not contain any secondary structure content. This criterion is satisfied in all the native crystal structures chosen. Optimization of these native structures, however, often reveals the presence of some secondary structure content.
5. The presence of no more than four additional loop residues on either side of the selected residue interval.

6. The minimum distance between any loop atom and any atom from a neutral ligand or organic ion in the environment has to be >4 Å. For a metal ion ligand, this cutoff is 6.5 Å.

Selection criteria enumerated above are quite similar to the filtering criteria we used in our previous study. Briefly, the goal is to focus on evaluating the ability of the sampling algorithm and protein/solvent energy model to yield accurate loop predictions, without considering additional issues such as to whether the potential function employed for a metal center is accurate. As will be seen below, despite the pH criterion (intended to minimize the number of "nonstandard" protonation states), ionization state issues have significant impact on a number of the loop predictions. We refer the reader to Ref. 1 for an extended discussion of the rationale behind these criteria.

RESULTS AND DISCUSSION

Overview

We begin by presenting a comparison between our previous method (denoted as Ref. 1) and the new method (denoted as *Modified Method*) for the 104 loop test suite described above. We assess our results solely by the "global" RMSD, obtained by superimposing the body of the protein and calculating the RMSD of the loop. An approximate comparison can also be made with the results in Table I; although the actual loops employed in other studies are different, the RMSDs to the native structure achieved by predictions should not be highly dependent upon the specific data set, as long as the loop lengths are equivalent.

To further clarify the impact of our new hydrophobic term, we also have run the new sampling algorithm without the hydrophobic term on the 13-residue loop data set. At a length of 13 residues, a wide range of alternative conformations exist for most loops, and the effects of incorporating the new hydrophobic term are dramatic in terms of the ability to select the native conformation from among this wide range of alternatives. The new sampling algorithm does produce significant reductions in sampling errors even for the old energy model; however, it is interesting to note that the improvement is qualitatively better in this dimension as well when the new hydrophobic potential is incorporated. This observation suggests that addition of the hydrophobic potential changes the nature of the potential surface, not just the relative depth of the various minima, reducing barriers to locating the native basin of attraction. Further investigation of this issue will be pursued elsewhere.

Finally, we focus on a few cases where we observe failure of our methodology and discuss the potential remedies. In a significant number of cases, the poor results can be eliminated by relatively straightforward improvements: expansion of the backbone rotamer library (to overcome a sampling error), more rigorous assignment of protonation states to correct the energy model, or extension of the hydrophobic packing term to intermolecular interactions. Larger data sets will be required to fully test and optimize these augmentations of the methodology, but the initial results are encouraging. A small fraction of cases continue to display substantial errors (greater than 2 Å RMSD), for reasons as yet unknown; further optimization of the energy model will be required if truly robust results are to be obtained on this length scale (one that is quite different from the 5–10 Å scale errors, which are quite common at a length of 11–13 residues in all previously published loop prediction efforts).

TABLE II. Comparison of Method in Ref. 1 and Our New Method

	Methods from Ref. 1		New Method	
	Mean	Median	Mean	Median
11 residue loops	2.03	1.31	1.00	0.62
12 residue loops	2.52	2.05	1.15	0.60
13 residue loops	4.13	3.51	1.25	0.76

The accuracies are measured by global backbone RMSDs in Å.

Overall Performance

Table II compares summarized results obtained using the sampling algorithms and energy functions defined in Ref. 1 with those arising from our improved approach. The detailed data for all cases can be obtained in the Supplementary Material. For our large data sets of 11- and 12-residue loops, the old methodology yields the mean global backbone RMSDs of 2.03 and 2.52 Å, respectively, which are close to the results reported in our previous study (employing a different, much smaller data set) at these lengths. Because of the paucity of data at these loop sizes in that work, we did not characterize these results as anything more than anecdotal. The new large data set results clearly show our previous methodology becomes problematic in general for such large loop sizes. Furthermore, the results for 13-residue loops exhibit a severe degradation, analogous to that seen in other work as indicated in Table I: the mean and median RMSD are 4.13 and 3.51 Å, respectively.

Our new methodology obtains uniformly excellent results for 11- to 13-residue loop data sets, with the average RMSD ranging from 1.00 to 1.25 Å. The median RMSD for each subset ranges from 0.52 to 0.76 Å. For 38 11-residue loops, only two cases (5.3%) have RMSD larger than 2 Å (one of the biggest outlier, exhibiting a 7.88 Å RMSD, can be fixed by an trivial optimization of the our scoring function; see discussion below). Twelve- and 13- residue loop data sets have 24 out of 31 (77.6%) and 27 out of 35 (77.1%) cases with RMSD below 2 Å. Furthermore, if we choose 1 Å as the threshold, there will be 63.2, 67.7, and 57.1% cases with RMSD smaller than 1 Å for 11-, 12-, and 13-residue data sets, respectively. These results clearly show a substantial improvement over our previous method. Even though it is difficult to make rigorous comparisons with the other published methods, our results unambiguously represent a breakthrough in this kind of calculation considering the fact that all those methods (including our previous method) generated very bad predictions for long

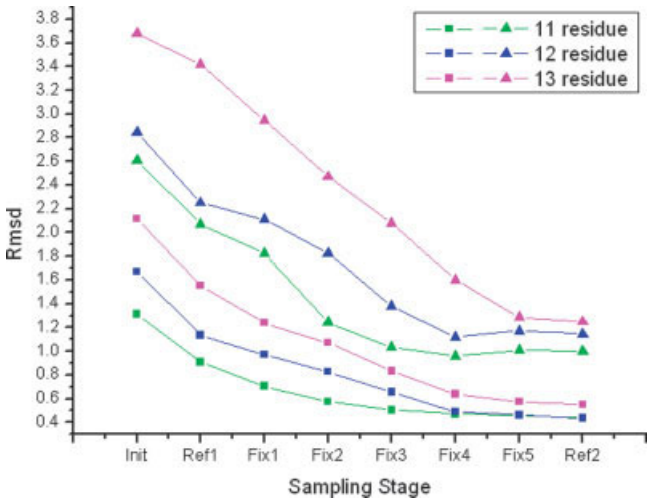

Fig. 3. Predicted and best-generated conformations with respect to sampling stages. The triangle and square represent for predicted and best-generated conformations, respectively.

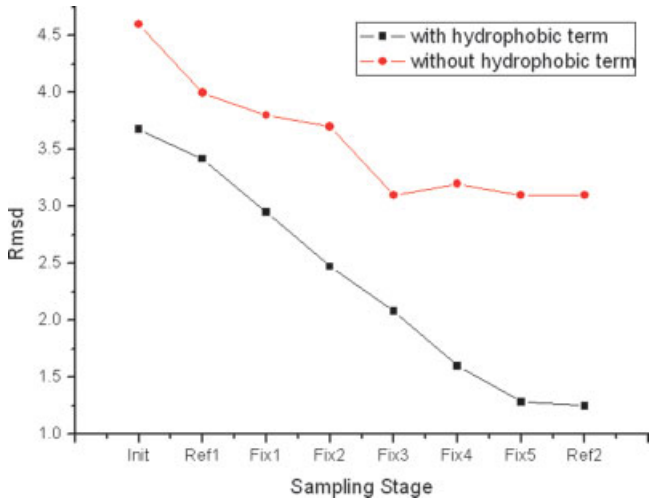

Fig. 4. Prediction with and without hydrophobic term for 13-residue sets using our new sampling algorithm.

loop targets.[1–5] The average RMSD now increases quantitatively (not qualitatively) compared to that for shorter loops, the number of large outliers has been drastically reduced, and the median RMSD is close to experimental accuracy. In contrast, all previous work yields results for loop lengths of 12 and above that is close to random noise, with average RMSDs in the range of 3.4–4.9 Å and a high fraction of RMSDs greater than 5 Å, which would be essentially useless for any sort of practical application of the model.

Sampling algorithm

Our new sampling algorithm plays an essential role in obtaining good results, addressing the exponential explosion of possible conformations with loop length by focusing on a series of small segments. Our sampling algorithm has solved this problem in some sense, generally eliminating the sampling errors (which is defined as the failure case (RMSD $>$ 2 Å) with higher energy than native structure) in our results. For our 104 targets, there are only two cases (1k7i in an 11-residue subset and 1p1m in a 13-residue subset) showing significant sampling errors (based on an analysis of the energy gaps) and a careful examination (see below) shows that these two cases can be attributed to factors other than the sampling algorithm, such as the deficiency of the scoring function and incompleteness of the backbone library. This further confirms the effectiveness of our sampling methodology.

Figure 3 shows the average RMSD of 11-, 12-, and 13-residue loops obtained from the various sampling stages. In each stage, our sampling algorithm generates many models and ranks them according to the energy function and then passes the lowest energy model(s) to the next stage. To get a sense of how well our sampling method performs at each stage, we also rank the scored models according to their RMSDs for each stage and show them in Figure 3. This is the prediction accuracy limit of our sampling algorithm. There is a clear tendency for both RMSDs to keep decreasing as the sampling stages are advanced. The initially predicted structures are usually not satisfactory, but our sampling algorithm continues to generate more and more accurate structures, and our energy function is able to select them out, abd thus keeps improving prediction accuracy.

1p1m represents a major sampling challenge for the methodology. We do not see the common tendency of decreasing energies and RMSDs as the sampling stage move on. There is a positive energy gap of around 30–40 kcal/mol after the *Fix2* stage, which does not further diminish as refined sampling is applied. We suspected that this happened because our backbone library limits the buildup procedure to generate a native-like structure. To test this point, we changed to a very large fine-grained library composed by DePristo et al.[22] This library is composed from 500 nonredundant protein structures. It has a 5° resolution (significantly higher than our default library) and could cover 99.99% of the (ϕ, ψ) combinations appearing in the native structure. There are 4264 rotamer states for Gly, 1207 for Pro, and the average 2666 states for the other residues. The prediction with this new library generates a 0.94 Å RMSD structure and 7.6 kcal/mol positive energy gap, as shown in Table III.

However, we should not simply take it for granted that the larger library is necessarily better than the smaller one overall. The number of candidates generated in the buildup stage is prespecified; it can only fluctuate within a certain range. The larger backbone library, which contains more possible conformation states, would cause more coarse "effective resolution," on the other hand, a smaller library would allow us to sample more densely on the high propensity region of Ramachandran plot. The balance between the "effective resolution" and library size needs to be carefully considered. In further work, we intend to test the DePristo et al. library on our entire test set, and also to increase the size of the test set to locate more difficult sampling cases. Ultimately, we will refine the backbone library to increase the robustness of the methodology; as

TABLE III. Comparison of Our Library and Rapper72 Library on 1p1m Case

	Init	Ref. 1	Fix1	Fix2	Fix3	Fix4	Fix5	Ref. 2
OUR	109.1/2.66	68.7/9.39	58.0/3.69	42.2/3.77	42.2/3.77	42.2/3.77	31.4/3.65	31.4/3.65
RAPPER72	103.9/3.79	69.9/3.62	67.2/3.15	57.0/2.82	54.5/2.95	50.3/1.74	7.6/0.94	7.7/0.94

The energy gap(kcal/mol) and RMSD(Å) are shown for each stage.

TABLE IV. Comparison of Three Results on 13-Residue Targets

			Method from Ref. 1		New Sampling		New Sampling + Hydrophobic	
PDB	RESLO	RESHI	dE	RMSD	dE	RMSD	dE	RMSD
1ojq	A:167	A:179	−0.02	6.36	−23.13	4.45	−16.25	4.06
1dys	A:290	A:302	16.38	1.96	−1.32	0.26	−6.98	0.28
1dpg	A:352	A:364	−7.25	10.09	−10.70	8.26	−22.18	1.27
1xyz	A:645	A:657	16.57	4.70	−5.28	0.36	−1.32	0.36
1eok	A:147	A:159	−2.54	0.52	−3.74	0.43	−20.47	0.37
1plm	A:327	A:339	14.60	9.09	6.23	7.25	31.42	3.65
1ock	A:43	A:55	−29.08	3.95	−35.00	3.94	−33.39	2.90
1hnj	A:191	A:203	−15.46	6.20	−26.08	3.21	−43.77	3.11
1iir	A:197	A:209	−7.05	0.18	−7.89	0.18	−21.29	0.21
1h4a	X:19	X:31	32.70	2.84	−1.37	0.27	−13.49	0.26
1o61	A:386	A:398	5.25	7.49	−5.90	6.48	−7.56	1.05
1arb	−:182	−:194	5.09	2.40	−5.82	2.55	−17.75	0.85
1bkp	A:51	A:63	6.88	0.34	1.13	0.83	−7.53	0.83
1f46	A:64	A:76	8.04	7.76	−11.06	0.71	−5.60	1.27
1hxh	A:87	A:99	4.18	0.52	0.08	0.42	4.64	0.81
1jp4	A:153	A:165	−30.73	3.20	−31.31	3.59	−26.02	3.43
1nln	A:26	A:38	11.27	2.76	−5.24	0.57	−13.36	0.71
1kbl	A:793	A:805	3.36	7.59	−44.21	5.84	−4.83	0.48
1l8a	A:691	A:703	4.60	0.35	0.80	0.23	−1.76	0.25
1a8d	−:155	−:167	41.18	2.74	28.95	3.29	−3.89	0.33
1bhe	−:121	−:133	8.05	3.51	−13.40	2.49	−15.42	2.45
1cnv	−:110	−:122	1.96	5.68	−5.79	6.06	−6.34	1.03
1mo9	A:107	A:119	−19.30	10.24	−32.93	9.56	−2.80	0.76
1gpi	A:308	A:320	7.22	8.28	2.33	7.00	5.29	0.70
2ptd	−:136	−:148	49.96	2.82	−16.98	0.34	−28.36	0.46
1lki	−:62	−:74	−2.84	0.69	−4.53	0.83	−15.69	0.36
1qqp	2:161	2:173	47.71	5.88	45.57	6.30	−13.69	0.38
1g8f	A:72	A:84	56.66	5.72	−12.57	1.44	−25.74	1.41
1qsl	A:389	A:401	−13.52	5.15	−20.75	4.31	−33.60	3.61
1d0c	A:280	A:292	29.94	2.60	20.59	5.87	−2.01	0.30
1krh	A:131	A:143	22.84	4.52	7.30	4.46	3.02	0.72
2hlc	A:91	A:103	4.48	4.03	−16.31	4.37	−2.74	3.28
1ako	−:203	−:215	−6.93	0.33	−18.67	1.29	−26.87	1.07
1ed8	A:67	A:79	62.09	3.41	−10.00	0.26	−20.81	0.26
1m8s	A:68	A:80	3.36	0.59	0.55	0.54	−0.53	0.45
		Average	9.42	4.13	−7.33	3.09	−11.93	1.25
		Standard deviation	22.31	2.96	17.39	2.77	14.13	1.20
		Median	5.09	3.51	−5.82	2.55	−13.36	0.76

New sampling refers to our new sampling algorithm with the unmodified OPLS and SGBNP energy function. New sampling + hydrophobic refers to the combination of our new sampling algorithm and new hydrophobic term.

suggested above, a large test suite (and one including longer loops) is required to do this rigorously.

Scoring Function

To clearly show the advantage of including our new hydrophobic term, we also have run our sampling algorithm on the 13-residue data set with the unmodified OPLS and SGBNP energy function. The result is shown in Figure 4 and Tables IV and V. The sole application of our new sampling algorithm could lead to some improvement compared to the algorithm of Ref. 1, by reducing the average RMSD from 4.13 to 3.09 Å. It is certainly an improvement, but does not have nearly as great impact as when combined with the hydrophobic term. Further investigations show there are a number of cases that converge to an erroneous structure with huge RMSD and negative

TABLE V. Six Best Improved Cases by the Inclusion of Hydrophobic Energy

		OPLS/SGBNP	Hydrophobic	Total energy	RMSD
1dpg	without Lipo	−46371.28	−83.98	−46455.26	8.26
	with Lipo	−46366.50	−162.50	−46529.00	1.27
1o61	without Lipo	−17027.42	−65.73	−17093.15	6.48
	with Lipo	−17017.50	−103.54	−17121.05	1.05
1kbl	without Lipo	−39388.35	−145.73	−39534.07	5.84
	with Lipo	−39354.43	−215.67	−39570.10	0.48
1cnv	without Lipo	−12871.88	−99.50	−12971.39	6.06
	with Lipo	−12865.07	−157.64	−13022.71	1.03
1mo9	without Lipo	−47805.86	−61.87	−47867.72	9.56
	with Lipo	−47767.41	−190.91	−47958.32	0.76
1gpi	without Lipo	−17895.14	−68.56	−17963.69	7.00
	with Lipo	−17809.93	−107.22	−17917.15	0.70

For each case, the best structures in the prediction experiments with and without hydrophobic energy (denoted by *withLipo* and *withoutLipo,* respectively) are generated and their energies (kcal/mol) and RMSDs (Å) are listed in the table. For every case, the bad structure has the lower OPLS/SGBNP energy, however, the good structure wins over after the hydrophobic energy is included.

energy gap. Many of these cases are fixed by the inclusion of hydrophobic term, as shown in Table V.

In our current implementation of hydrophobic term, we do not include the contributions from the symmetry copies of other protein molecules in the unit cell. Such contributions are of course not relevant in a realistic biological application (i.e., in solution), but are relevant to comparisons with crystallographic data as in the present article. In our prediction experiments, we notice some loop targets are "sticking out" from the protein molecule to which they are attached, and do not have close contact with the body of that protein. We speculate that the major forces that stabilize a loop with this sort of topology should come from the symmetric copies. 1k7i in the 11-residue subset represents an example of this type. In our current implementation (i.e., without hydrophobic contributions from symmetric copies), we get a prediction accuracy of RMSD 7.8 Å and a positive energy gap of 7 kcal/mol. This predicted structure is very close to the one generated with our previous protocol (i.e., without the hydrophobic term and new sampling algorithm). After including the hydrophobic contribution from symmetric copies, we get an almost perfect prediction with RMSD 0.42 Å and energy gap of −4.1 kcal/mol, as shown in Figure 5. This case suggests it might be desirable to include hydrophobic term from symmetric copies into our scoring function in the future when investigating loop prediction in the crystal environment, because this contribution may play an essential role for the stabilization of the loop. We calculated the self-terms (i.e., those from the central protein molecule) and symmetric (i.e., symmetric copies atoms) contributions to the hydrophobic energy according our definition for all the 104 targets. The results (data not shown) show that for the great majority of the targets the symmetric contribution is much smaller than the self contributions, and only in a few cases the symmetric contributions are comparable to the asymmetric contributions. For 1k7i, the symmetric and asymmetric contributions are −54.51 and −49.51 kcal/mol. Our preliminary tests also show that including symmetric contributions does not significantly improve the prediction accuracy from the statistical point of view, except that it is helpful for very few cases. Therefore, we do not explore further the inclusion the symmetric hydrophobic contributions in this work. We will leave this task, together with the optimization of the coefficients of hydrophobic terms, to subsequent publications. As noted above, inclusion of interactions with symmetry copies is irrelevant for actual biological applications which do not take place in the crystal.

Our implementation of hydrophobic term uses the same functional form as in ChemScore. Because it does not have the first-order derivative, this term is only used in the screening and scoring, and not in the minimization of the structures. We aim to develop an analytical form in the future optimization to make it suitable for wider applications such as minimization and molecular dynamics. Because this term is in general significantly more slowly varying than van der Waals terms and smaller than electrostatic and solvation terms, we would expect that including the ChemScore term in the gradient would not change the local minimization significantly. In the context of structure predictions, a conformational sampling algorithm capable of generating a variety of candidate structures is much more critical than the minimization. Thus, a hydrophobic term with analytical gradient would not change the loop prediction results significantly.

Protonation State Assignment

As we discussed in our previous study, the pH has a strong effect on the protonation states of ionizable groups, and using standard protonation states (Asp, Glu ionized; Arg, Lys protonated; His neutral) at neutral pH is not necessarily the optimal assignment. The existence of alternative protonation states can have a significant effect on the accuracy of loop prediction. Recently, a general algorithm has been developed in our lab to assign the protonation states for ionizable residues, together with the positioning the polar hydrogens (i.e., hydrogens on the —OH group of carboxylic acid/C-terminal end groups and $NH3^+$ groups on Lysine side chains/N-terminal end groups) and the side

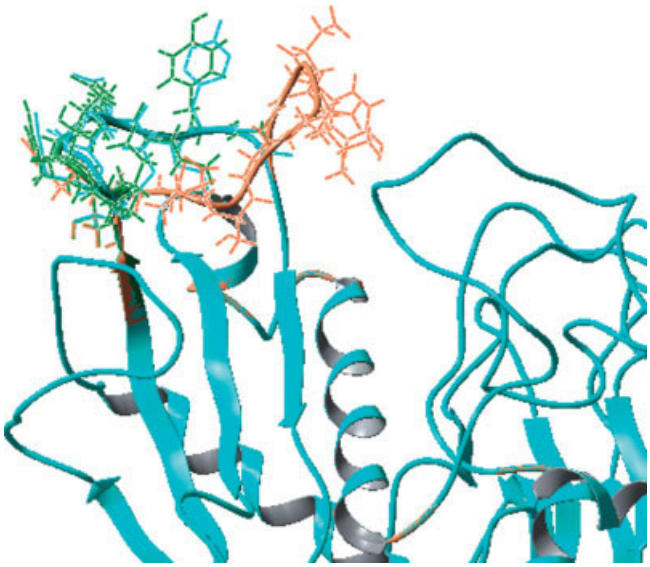

Fig. 5. Hydrophobic interactions from symmetric copies stabilize the loop in 1k7i. \the cyan represents for native structure. The green and brown represent for the two predicted structures with and without hydrophobic interactions from symmetric copies. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

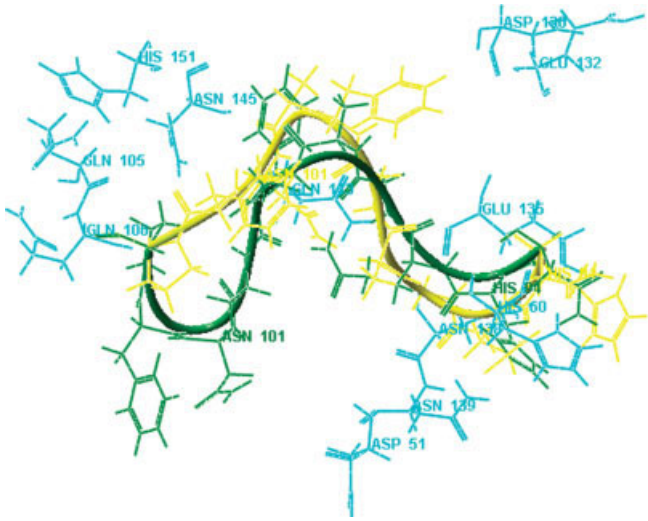

Fig. 6. Protonation stage assignments affect prediction for 1edt. All the residues changed by assignment algorithm are labeled out and colored by cyan. The green and yellow represent for two predictions with assignment structure and original PDB structure, respectively. The native structure is not shown because it is almost identical with the green prediction structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

chain $\chi 2$ flipping of the amide groups in Gln/Asn/His.[23] This method partitions all polar residues into disjoint clusters based on local hydrogen bonding network and then each cluster states set can be optimized independently to reduce the sampling space. In each cluster, all the possible states (protonation states, polar hydrogen positions, and flipping of $\chi 2$) are enumerated and evaluated with the OPLS-AA force field and SGBNP implicit solvent model. Finally, all the optimal state sets in each cluster are combined to give the final assignment.

In our loop prediction experiments, we have checked some bad predictions and found that protonation state misassignment may be responsible for the failure. We have run our assignment algorithm on a majority of the substantive prediction failure cases, and found that the assignment algorithm gave alternative protonation states in the target loop region or its neighbor region for several cases (1eur and 1edt in an 11-residue set, 1hnj in a 13-residue set). Rerunning our prediction with the structure generated by the assignment algorithm showed substantial improvements (1eur from 4.58 to 1.74 Å, 1edt from 5.94 to 0.28 Å and 1hnj from 8.28 to 3.11 Å). We have included these new predictions in our results. Figure 6 shows the 1edt target (residue 93 to 103) and its neighbor environment. Within 10 Å range of the target loop, there are a total of 14 residues whose protonation states needed to be changed or $\chi 2$ angles needed to be flipped based on our assignment algorithm. Also, His 94 in this loop region should be protonated on both nitrogen atoms. Using the new assignment structure yields a 0.28 Å RMSD prediction while the standard assignment predicts a structure with 5.94 Å RMSD. We do not see any single hydrogen bond or salt bridge responsible for the previous prediction failure; however, it is reasonable to attribute the success using the new assignment structures to a more correct description of the electrostatic environment.

TABLE VI. Computational Costs for 11–13 Residue Data Sets

	CPU time (day)			
	Mean	Median	Min	Max
11 residue loops	12.3	11.7	5.0	25.4
12 residue loops	19.1	15.6	8.4	43.9
13 residue loops	31.4	22.5	12.3	114.6

The CPU time refers to the cumulative time counted on a single processor.

Computational Costs

Table VI shows the various statistics of computational time for our 11-13 residue data sets. Calculations are conducted on a cluster of Intel or AMD processors in the range of 1.4 GHz or 900 MHz. We randomly distribute the computational jobs into the various processors. Even there is some difference in the performance of the various processors, but this would not affect the qualitative analysis of our methodology. The average CPU time for 11-residue loop target is about 12 days; a 12-residue loop takes about 19 days and a 13-residue loop takes about 31 days, which is 2.5 times of the 11-residue. On the other hand, because most of the calculations are highly parallel, a typical 13-residue loop prediction can be actually finished in 2–3 days on a cluster of 16 processors. We are seeking a variety of efforts to speed up the algorithm to make it feasible to carry out longer loop prediction.

DISCUSSION

The success of the new loop refinement algorithm that we have developed is straightforward to understand. By focusing sampling effort on loop segments, rather than the entire loop, the phase space of the segment can be investigated in more detail, and relatively deep, but narrow, basins of attraction can be located and converged to the appropriate total energy. It remains to be seen whether the robustness displayed for the present test set (essentially no sampling errors due to the algorithm observed in more than 100 test cases) persists as the loop length increases; at some point, augmentations of the present algorithm may have to be developed. There are a number of strategies that can be deployed as loop length increases, and we do not believe that there will be fundamental problems in extended our sampling strategy to loops in the 15–20-residue range.

The dramatic improvement in the energy model that has been obtained by adding a hydrophobic term, on the other hand, is rather surprising (although undeniable in efficaciousness—the new model contains in essence only one adjustable parameter, and that parameter has not actually been optimized to obtain the present set of results, as discussed above), and has significant implications for the design of approximate continuum solvation models. This is a subject that has received a great deal of attention over the past decade in the literature; however, the focus[24–26] has overwhelmingly been on reproducing the results of precise (in the computational, not physical, sense) numerical Poisson-Boltzmann calculations, typically augmented by surface area term to model the hydrophobic effect (yielding a "PB/SA" model, or a GB/SA approach fit to such a model). This is a worthy goal, but it ignores the question of how well the PB/SA model itself describes physical reality. In contrast, our view has been that all continuum models need to be compared with experimental data, given the gross approximations made in any such model (including numerically "accurate" PB/SA calculations) compared to a realistic description of aqueous solvation. We have chosen loop and single side chain prediction as our initial testing ground; if one cannot predict localized structures in the well-defined, exact native protein environment, it is difficult to see how much more sensitive quantities, such as ligand binding affinities, are going to be evaluated to the required 0.5–1.0 kcal/mol level needed for major impact on problems such as lead optimization in drug discovery.

The key initial issue in improving continuum solvation models is to identify the leading source of error. The present results unambiguously demonstrate that, on lengths scales relevant to long-loop prediction, the error is completely dominated by inaccurate descriptions of the hydrophobic effect provided, for example, by conventional surface area models. Such models are typically fit to small molecule experimental solvation free energy data, which does not guarantee that they will work well when describing water molecules in the constricted environment of a hydrophobic cavity in a large, complex protein structure. The adequacy of the functional form, as well as the parametrization, must be called into question.

There are undoubtedly errors in the present model in its description of hydrogen bonding and electrostatic interaction of polar and charged groups in solution. Some of these errors may arise from inaccuracies in our surface Generalized Born treatment, while others are intrinsic to any sort of continuum model; for example, we have recently shown that all such models display large (3–5 kcal/mol) errors in modeling a single bridging water between two oppositely charged groups.[27] There is also a general tendency of the current generation of continuum models to overpredict formation of solvent exposed salt bridges. However, these errors apparently have an effect primarily on shorter length scales than those relevant to obtaining the correct loop backbone conformations, for example, in the prediction of side-chain rotamer states. A remarkable aspect of the results in the present article is how well continuum electrostatics performs in loop prediction, despite the (arguable) crudeness of the present SGB model, once complemented by an improved description of hydrophobicity.

We now briefly consider the question of why models of hydrophobicity based on optimization to small molecule solvation data might become very inaccurate for larger structures such as proteins. As an extreme case, consider the placement of a single water into a cavity that is hydrophobic on all sides (note that such cavities are not manifested for molecules of the size typically used to parametrize continuum hydrophobic models). Because the water cannot form any hydrogen bonds, the free energy cost of such a transfer is extremely large. However, in a continuum model, there is no energetic penalty applied to the dielectric 80 volume assigned to the cavity. Elimination of the cavity would be rewarded by an increase in van der Waals energy of the solute, but there is no a priori reason to believe that this increase would properly measure the free energy gain as compared to having a void (which requires excluding solvent, a very unfavorable process energetically), or allowing a water molecule to occupy the cavity while unable to make any hydrogen bonds.

A simple model for "docking" a loop of significant length into the body of the protein is that the process is dominated by packing of hydrophobic side chains into hydrophobic pockets of the remainder of the protein. This process is then very similar to docking of a ligand into a protein active site. Based on the argument given above, it is plausible to hypothesize that current continuum models underestimate the gain in free energy available from formation of a hydrophobic core. False positives can then be generated by small gains in electrostatics, at the expense of proper packing of hydrophobic side chains into the appropriate pockets, being selected. In Figure 7, we compare a false positive loop structure for 1mo9 (generated as lowest energy without the use of the new ChemScore-based hydrophobic term) with the native-like (0.76 Å) structure produced when the new hydrophobic term is employed in the optimization. The qualitatively superior

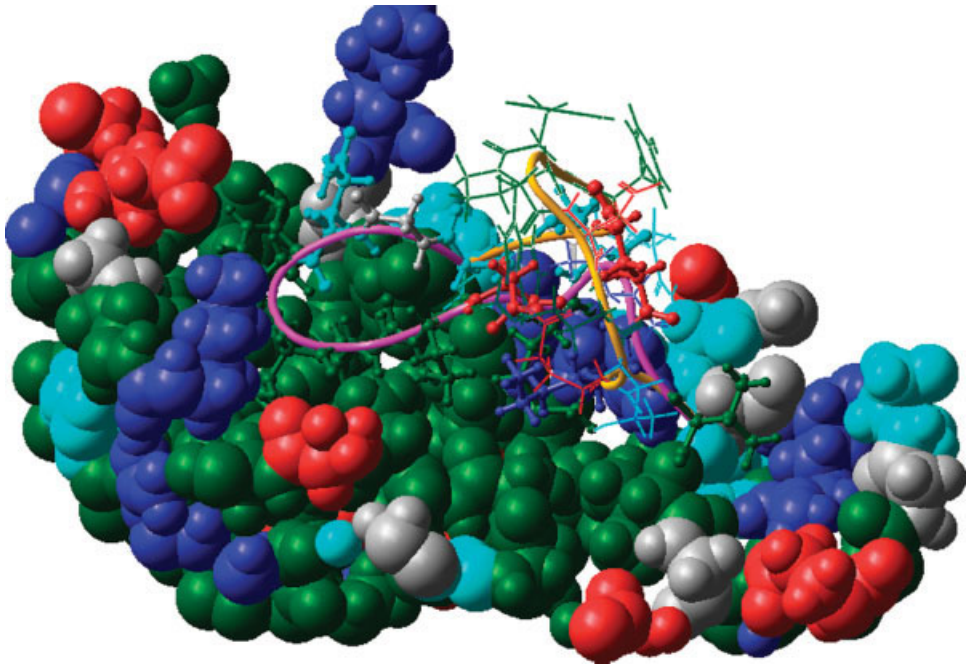

Fig. 7. Hydrophobic packing between the loop hydrophobic side chains and the hydrophobic pocket of the protein body for 1mo9 case. The green color represents for hydrophobic residues, blue for positive charged residues, red for negative charged residues, cyan for polar uncharged residues, and gray for Gly. All the atoms within 10 Å of the loop target are shown. The two loops, correctly predicted structure (0.76 Å RMSD with native structure) and false positive without hydrophobic term (9.56 Å RMSD), are represented by purple and yellow cartoons. Hydrophobic contacts are obvious in (nearly) native structure, but they are absent in false positive structure. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

filling of hydrophobic pockets by the native-like structure is apparent.

In this picture, we expect that there will still be occasional errors in the prediction of polar and charged side-chain conformations, even if the backbone conformation is greatly improved. These errors typically are indicative of the availability of alternative hydrogen bonding structures at the native backbone conformation, or positioning of the side chain in solution as opposed to forming a hydrogen bond or salt bridge. Accurately and robustly ranking these alternatives is a very challenging problem (and one that we do not claim to have solved in the present article), but it is entirely plausible that one can achieve reliable backbone prediction (to within the ~1–2 Å indicated in Table II) without perfect specification of the side-chain hydrogen bonding patterns. The most powerful argument along these lines arises from the observation that NMR and X-ray crystallographic loop structures of the same protein are very similar (often within ~1 Å RMSD, and rarely more than 2 Å RMSD apart), despite the fact that in many cases (where crystal packing is important) the side chains of the loops form very different hydrogen bonding patterns. Put another way, the hydrogen bonding patterns of the polar and charged groups are opportunistic and many alternative conformations likely yield similar free energies; as long as all of the relevant hydrogen bonds can be satisfied, and one is not burying a net charge (in the absence of a suitable countercharge at close range), the basic underlying backbone structure can remain intact. In contrast, formation of the hydrophobic core as discussed above is critical and specific, within the constraints of satisfying the required hydrogen bonds in some fashion.

Having said this, it is still remarkable how well the composite PLOP/ChemScore energy function works at a quantitative level, particularly as no parameter optimization has been carried out. Some of this is no doubt due to the fact that structural prediction is less demanding than prediction of relative free energies of the alternative structures. If the structures containing the properly formed hydrophobic core are now overstabilized compared to the alternatives, this error will not be manifested in our results. Application of the scoring function proposed here to ligand binding will enable further testing and refinement of these ideas.

Finally, we should note that alternative approaches to the development of improved continuum hydrophobic models have been developed, most prominently by the Levy group in a series of articles over the past several years.[13,28–31] The most recent of these articles demonstrates impressive results of their model AGBNP in improving the prediction of the structure of small peptides, compared to standard continuum approaches.[29] It is quite possible that this model (which may have other advantages as compared to our ChemScore implementation) will succeed equally well, or better, in improving long loop prediction. This model has recently been implemented in

PLOP, so we will be able to test this hypothesis in the near future.

CONCLUSION

We have developed an improved sampling algorithm and energy model for protein loop prediction, the combination of which has yielded the first methodology capable of achieving good results for the prediction of loop backbone conformations of length greater than ~10–12 residues. The methodology appears to be both accurate and robust, with a very small number of outliers in evidence, some of which can be eliminated by straightforward improvements in the sampling or initial structure preparation. The computational effort required, while nontrivial, is tractable using relatively small clusters of personal computers, and can likely be reduced via optimization of the details of the algorithms.

The successful synthesis of a conventional continuum solvation model with an empirical hydrophobic packing term has significant implications for the development of continuum solvation models for biomolecular simulation. It is clear from the present results that the GB/SA approach used in the present article (and, presumably, other PB/SA and GB/SA approaches as well, because there is no reason to believe the latter are superior with regard to treatment of the hydrophobic effect) underestimates the free energy gained by filling predominantly hydrophobic pockets with hydrophobic side chain groups, that is, forming a hydrophobic core. The rationalization for this observation has been discussed above. We can hypothesize that this flaw is at present the leading error in all continuum solvation models, and that efforts to improve such models that ignore the problem cannot improve results beyond a certain point, at least if one is going to compare with experiment. Of course, a comparison between models (e.g., GB and PB) that contain the flaw in roughly equal measure will not reveal any difficulty. That is why detailed comparison with the experiment, as has been carried out in the present article, is necessary to validate any claims of improved accuracy.

Preliminary results indicate that the hydrophobic packing term, although essential for improving the loop backbone, has much less significant effects on side chain prediction accuracy. This observation, while disappointing, is not surprising in view of the analysis above. The specificity of loop backbone conformations are driven by formation of a hydrophobic core on a length scale of tens of Angstroms; on this length scale, errors in quantitation of the hydrophobic effect by the continuum solvation model are disastrous. However, side-chain conformations are typically determined by comparison of the free energy of alternative hydrogen bonding patterns (either by comparing two or more such alternative conformations with each other, or with a third alternative of placing the side chain in solution without making any hydrogen bonds to other protein groups). Getting such comparisons right requires accurate calibration of the delicate balance of electrostatic forces; achieving this in a robust fashion is still beyond the capabilities of current continuum solvation models, although improvements can certainly be made by fitting to experimental side chain structural data, as we have discussed elsewhere.

Future improvement of the overall energy model will require tackling the above problem in a serious fashion, as well as continued refinement of the hydrophobic component by examining a much larger set of test data. Although the present results are far from a complete solution, they do represent measurable progress, and hopefully will constitute a starting point for the development of models that are truly capable of accuracy localized structural prediction and, ultimately, refinement of homology models to atomic resolution.

ACKNOWLEDGMENTS

R.A.F. thanks Barry Honig for useful discussions.

REFERENCES

1. Jacobson MP, Pincus DL, Rapp CS, Day TJF, Honig B, Shaw DW, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins Struct Funct Bioinformat 2004;55:351–367.
2. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
3. de Bakker PIW, Depristo MA, Burke DF, Blundell TL. *Ab initio* construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the Generalized Born solvation model. Proteins Struct Funct Bioinformat 2002;51:21–40.
4. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
5. Rohl CA, Strauss CE, Chivian D, Baker D. Modeling structurally variable regions in homologous proteins with rosetta. Proteins Struct Funct Bioinformat 2004;55:656–677.
6. Monnigmann M, Floudas CA. Protein loop structure prediction with flexible stem geometries. Proteins Struct Funct Bioinformat 2005;61:748–762.
7. Hartigan JA. Clustering algorithms. New York: Wiley; 1975.
8. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Appl Stat 1979;28:100–108.
9. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
10. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 2001;105:517–529.
11. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002;105:11673–11680.
12. Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.
13. Gallicchio E, Zhang LY, Levy RM. The SGB/NP hydration free energy model based on the surface generalized Born solvent reaction field and novel nonpolar hydration free energy estimator. J Comput Chem 2002;23:517–529.
14. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11:425–445.
15. Jacobson MP, Friesner RA, Xiang ZX, Honig B. On the role of the crystal environment in determining protein side-chain conformations. J Mol Biol 2002;320:597–608.
16. Xiang ZX, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
17. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
18. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD.

Improved Protein–ligand docking using GOLD. Proteins Struct Funct Bioinformat 2003;52:609–623.
19. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47:1739–1749.
20. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 2004;47:1750–1759.
21. Wang GL, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
22. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins Struct Funct Bioinformat 2003;51:41–55.
23. Xin L. Advances in high resolution protein structure modeling. New York: Columbia University; 2005.
24. Lee MS, Salsbury FR Jr, Brooks CL. Novel generalized Born Methods. J Chem Phys 2002;116:10606–10614.
25. Romanov AN, Jabin SN, Martynov YB, Sulimov AV,Grigoriev FV, Sulimov VB. Surface Generalized Born Method: a simple, fast and precise implicit solvent model beyond the Coulomb approxiMation. J Phys Chem A 2004;108:9323–9327.
26. Wojciechowski M, Lesyng B. Generalized Born Model: analysis, refinement,and applications to proteins. J Phys Chem B 2004;108: 18368–18376.
27. Yu Z, Jacobson MP, Rapp CS, Friesner RA. First shell solvation of ion pairs: correction of systematic errors in implicit solvent models. J Phys Chem B 2004;108:6643–6654.
28. Levy RM, Zhang LY, Gallicchio E, AK F. On the nonpolar hydration free energy of proteins: surface area and continuum solvent models for the solute–solvent interaction energy. J Am Chem Soc 2003;125:9523–9530.
29. Felts AK, Harano Y, Gallicchio E, RM L. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. Proteins Struct Funct Bioinformat 2004;56:310–321.
30. Gallicchio E, RM L. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. J Comp Chem 2004;25:479–499.
31. Banks JL, Beard HS, Cao Y, Cho AE, Damm W, Farid R, Felts AK, Halgren TA, Mainz DT, Maple JR, Murphy R, Philipp DM, Repasky MP, Zhang LY, Berne BJ, Friesner RA, Gallicchio E, Levy RM. Integrated modeling program, applied chemical theory (IMPACT). J Comp Chem 2005;26:1752–1780.