PROTEINS: Structure, Function, and Bioinformatics 55:351–367 (2004)

A Hierarchical Approach to All-Atom Protein Loop Prediction

Matthew P. Jacobson,[1]* David L. Pincus,[2] Chaya S. Rapp,[3] Tyler J.F. Day,[4] Barry Honig,[5] David E. Shaw,[4] and Richard A. Friesner[2]

[1]*Department of Pharmaceutical Chemistry, University of California, San Francisco, California*
[2]*Department of Chemistry, Columbia University, New York, New York*
[3]*Department of Chemistry, Stern College, Yeshiva University, New York, New York*
[4]*Schrödinger, Inc., New York, New York*
[5]*Howard Hughes Medical Institute and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York*

ABSTRACT **The application of all-atom force fields (and explicit or implicit solvent models) to protein homology-modeling tasks such as side-chain and loop prediction remains challenging both because of the expense of the individual energy calculations and because of the difficulty of sampling the rugged all-atom energy surface. Here we address this challenge for the problem of loop prediction through the development of numerous new algorithms, with an emphasis on multiscale and hierarchical techniques. As a first step in evaluating the performance of our loop prediction algorithm, we have applied it to the problem of reconstructing loops in native structures; we also explicitly include crystal packing to provide a fair comparison with crystal structures. In brief, large numbers of loops are generated by using a dihedral angle-based buildup procedure followed by iterative cycles of clustering, side-chain optimization, and complete energy minimization of selected loop structures. We evaluate this method by using the largest test set yet used for validation of a loop prediction method, with a total of 833 loops ranging from 4 to 12 residues in length. Average/median backbone root-mean-square deviations (RMSDs) to the native structures (superimposing the body of the protein, not the loop itself) are 0.42/0.24 Å for 5 residue loops, 1.00/0.44 Å for 8 residue loops, and 2.47/1.83 Å for 11 residue loops. Median RMSDs are substantially lower than the averages because of a small number of outliers; the causes of these failures are examined in some detail, and many can be attributed to errors in assignment of protonation states of titratable residues, omission of ligands from the simulation, and, in a few cases, probable errors in the experimentally determined structures. When these obvious problems in the data sets are filtered out, average RMSDs to the native structures improve to 0.43 Å for 5 residue loops, 0.84 Å for 8 residue loops, and 1.63 Å for 11 residue loops. In the vast majority of cases, the method locates energy minima that are lower than or equal to that of the minimized native loop, thus indicating that sampling rarely limits prediction accuracy. The overall results are, to our knowledge, the best reported to date, and we attribute this success to the combination of an accurate all-atom energy function, efficient methods for loop buildup and side-chain optimization, and, especially for the longer loops, the hierarchical refinement protocol. Proteins 2004;55:351–367.** © 2004 Wiley-Liss, Inc.

Key words: generalized born solvent model; loop prediction; OPLS; all-atom force field; conformational sampling

INTRODUCTION

The problem of refining protein structures to high resolution, given a lower resolution initial structure, is becoming increasingly relevant to practical computational biology and structure-based drug design projects. Structural genomics efforts are increasing the number of structures in the Protein Data Bank at a rapid rate; however, the number of sequences is growing even more quickly. For a given sequence, one can very often find a reasonable template in the PDB from which to begin the process of constructing a homology model. However, once such a model is assembled (using, e.g., various alignment and model-building tools), it has proved to be very difficult to refine the model so that its structure moves toward the target to the point where it represents a significant improvement over the original model built from the template. Indeed, most frequently, attempts at refinement lead to degradation, rather than improvement, of the model compared to the actual target structure.

In principle, one should be able to solve the refinement problem by using an accurate molecular mechanics energy function (including solvation effects as well as interactions between atoms of the protein) and locating the global

Grant sponsor: National Science Foundation; Grant number: 0302445; Grant sponsor: National Institutes of Health; Grant numbers: GM-52018 and GM-30518.

*Correspondence to: Matthew P. Jacobson, Department of Pharmaceutical Chemistry, University of California, San Francisco, Box 2240, San Francisco, CA 94143–2240. E-mail: matt@cgl.ucsf.edu

Received 25 April 2003; Accepted 30 July 2003

Published online 5 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.10613

energy minimum of this function. In practice, a strategy along these lines has rarely been applied to the refinement problem because of the difficulties in achieving high accuracy in the energy function and the computational demands of sampling such a function. If a large number of explicit water molecules are used to describe solvation of the protein, standard simulation algorithms, such as molecular dynamics and Monte Carlo techniques, must be used to generate new configurations; these techniques require a large number of steps (and hence a large amount of computational effort) to traverse substantial regions of conformational space, which is required if refinement is to be successful for all but the simplest of cases. Continuum solvation methods offer more flexibility with regard to sampling algorithms, and hence, more progress has been made in using such methods for structural prediction, but the achievement of extensive and robust sampling is still a challenging endeavor. As a result, studies of this type are typically anecdotal, focusing on a small number of test cases, compared with the large data sets that are typically used to evaluate performance in less computationally intensive approaches to protein structure prediction.

Over the past several years, we have been developing a new approach that is aimed at reducing the computational effort for atomistic, continuum solvent-based predictions by several orders of magnitude, as well as improving the accuracy of the models themselves. The basic idea of the approach is to conduct an extensive search of conformational space using buildup techniques but to use hierarchical screening and clustering technology to enable a large conformational space to be searched in this fashion relatively quickly. Hierarchical screening methods ensure that the vast majority of proposed conformations can be quickly rejected because of, for example, steric overlap (although other, more subtle screens are also necessary if efficiency is to be achieved). Clustering methods reduce the number of conformations that need to be examined at a given level of the hierarchical screening procedure, by eliminating the need to score any structures in the cluster unless a representative member of the cluster achieves a reasonable score. The combination of hierarchical screening and clustering methods results in dramatic reductions in the scaling of the cost of a buildup algorithm with system size, as well as a similar diminishment in the prefactor. At the same time, the key advantages of a buildup methodology—exhaustive coverage of phase space, eliminating dependence on the starting configuration and topology of the energy landscape—can be preserved.

In the present article, we describe the computational methodology as applied to loops and focus primarily on an extensive evaluation of the ability of our current implementation to reconstruct loops in native structures using a large data set of loops of varying lengths. These calculations allow direct comparison of our results with others in the literature, as well as a statistically reliable assessment as to how well the energy function performs in this context. The calculations are performed in the crystal environment to obtain a more realistic comparison with X-ray crystallographic data. Although effects of the crystal environment on loop geometry are typically much smaller than on side-chain orientation (a subject we have discussed previously[1]), individual cases can manifest nontrivial differences, and obtaining an understanding of these differences is a useful endeavor in its own right.

Overview of Methodology

Details of the loop prediction methodology are discussed in Materials and Methods. However, because our algorithm includes several novel components, we provide here a brief overview of salient features.

It is not our intention to exhaustively review previous literature on loop prediction, which is quite extensive, except to draw a few critical distinctions among previous techniques that are relevant to define our own contribution. The first relevant distinction is between algorithms that attempt to identify an adequate loop fragment (or conformational class[2,3]) from databases compiled from large numbers of solved structures in the Protein Data Bank[4,5] and algorithms that build loop conformations "ab initio."[6,7,8–17]. Among the various approaches to ab initio modeling are Monte Carlo Simulated Annealing,[18] molecular dynamics simulated annealing,[9,15] "random tweak,"[7,8] and various dihedral angle sampling buildup methods,[6,10,14] sometimes supplemented with analytical loop closure methods.[17,19–21] Our sampling algorithm is most closely related to the ab initio buildup procedures, although the detailed algorithm is entirely new. For loops of up to 12 residues, we have found that sampling does not limit the accuracy of our method, with very few exceptions, because we succeed at creating loops with lower energies than the (minimized) native structure. However, there is no impediment to using a database of loop fragments instead of or in addition to our ab initio buildup algorithm, and we have implemented such a capability for certain homology-modeling applications. Only the ab initio results are presented here.

The other critical distinction among loop prediction algorithms is the method used to find the best [lowest root-mean-square deviation (RMSD) to native] candidate loop among the many possible structures, whether generated ab initio or taken from the PDB. The scoring schemes that have been used vary widely. At one extreme are methods in which energetics play little or no role, and scoring is accomplished by statistics-based functions (e.g., based on loop sequence[2] and stem matching,[2,5] or contact map representations[10]). Fiser et al.[9] used a mixture of molecular mechanics force-field terms and statistically derived terms for scoring, whereas Xiang et al.[8] used a scoring function that included steric interactions, hydrophobicity, torsional energy, and hydrogen bonding, but ignored many electrostatic effects. One of the biggest challenges for a physics-based loop-scoring scheme is the treatment of solvent. Some studies have used purely gas-phase energetics,[18] whereas the early study of Moult and James[6] used an image-charge method to represent charge screening, and several studies have used simple distance-dependent dielectric models.[13,14] Smith and Honig[22] investigated the usefulness of Finite Difference

Poisson–Boltzmann method on three loops and compared with gas phase and an Atomic Solvation Parameter treatment. Recently, Rapp and Friesner[15] used the AMBER force field with a Generalized Born solvent model but only presented results for two loops in ribonuclease, and de Bakker et al.[23] used a similar energy function on a large database of loop decoys. The energy function used here is similar (OPLS all-atom force field[24–26] and Surface Generalized Born model of solvation[27,28]), but the sampling algorithms used are entirely new and make it possible to apply a realistic, well-validated solvent model to a large test set of loops, with modest computational expense.

MATERIALS AND METHODS

The loop prediction algorithm consists of the following basic components:

1. Loop buildup, to generate many (generally thousands of) loop candidates
2. Clustering, to reduce redundancy and select representative candidates
3. Side-chain optimization on the selected candidate loops
4. Complete energy minimization of side-chain optimized loops.

All energetic evaluations are performed by using an all-atom energy model based on the OPLS-AA force field and the Surface Generalized Born implicit solvent model.

As discussed in detail below, the loop optimization algorithm is applied hierarchically and iteratively. The initial optimization stage consists of unconstrained loop generation to coarsely sample the conformational space. Refinement stages follow, in which constrained loop generation is used to more finely sample the conformational space surrounding selected low-energy loop conformations from the initial optimization. The refinement stage is, of course, optional, but it significantly improves accuracy for longer loops.

Table I presents illustrative results for one loop (3pte residues 78–86) at each step of the optimization procedure and is meant to provide a concrete example to accompany the remainder of this Materials and Methods section.

TABLE I. Details of Hierarchical Loop Prediction for 3pte, Residues 78–86

		Stage 1	Stage 2	Stage 3
Buildup	#	13044	108347	57500
	best	0.38	0.87	0.48
	worst	5.08	4.46	2.47
Screened	#	1336	15835	8539
	best	0.38	1.06	0.49
	worst	5.08	4.09	2.68
Clustered	#	38	38	38
	best	1.34	1.65	0.78
	worst	4.49	3.51	2.38
Scored	#	38	38	38
	best	1.39	0.60	0.22
	worst	4.54	3.63	2.57
Lowest E loop	energy	−14559.5	−14575.35	−14609
	RMSD	2.01	0.60	0.22

"Best" and "worst" refer to the smallest and largest backbone RMSDs obtained at various stages during the algorithm, whereas "#" represents the total number of loops retained at that point. Stages 1, 2, and 3 refer to the initial loop generation and two refinement stages, respectively. As described in Materials and Methods, multiple loop predictions are performed at each stage, and the predictions detailed here are merely representative (they correspond specifically to the cases shown in Fig. 4). Within each stage, the results are detailed for each individual step in the algorithm: buildup, the ab initio dihedral angle sampling; screened, the conformations retained after post-buildup screens; clustered, the conformations retained by the clustering algorithm; and scored, the cluster representatives after side-chain addition and energy minimization. The energy and backbone RMSDs are also reported for the lowest energy loop at each stage.

Loop Buildup Algorithm

The buildup algorithm developed for this work bears some similarity to the early CONGEN method of Bruccoleri and Karplus[14] and the more recent work of DePristo et al.,[17] although the details of the methods are entirely different. The similarity resides in the fact that these methods all use dihedral angle buildup procedure as the primary means of generating loop conformations for energetic scoring. As pointed out by Bruccoleri and Karplus, the disadvantage of a naive implementation of such an algorithm is exponential scaling of the computational expense with loop length.* Our solution to this problem is a hierarchical algorithm, which is described below.

The cornerstone of the loop-sampling methodology is dihedral angle search. In analogy to algorithms for side-chain packing, we have developed "rotamer libraries" for backbone dihedral angles (i.e., discretized versions of the well-known Ramachandran plots). Gly and Pro were considered separately from other residues because of well-known differences in the distribution of their backbone dihedral angles. To obtain the dihedral angle libraries, we have used a large (>500 structures), nonredundant database of high-resolution (<2.2 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 extremely high resolution of the backbone library was chosen to ensure that discretization error does not fundamentally limit the achievable accuracy. However, in practice it is not possible to sample the backbone dihedral angles for a loop of any nontrivial length using such a large

*We want to clarify that nonexponential scaling for other algorithms should not necessarily be considered an advantage, because the conformation space available to the loop can generally be expected to grow exponentially with loop length. Algorithms that scale linearly with loop length, if not designed carefully, may simple not provide an adequate sampling of the available conformational space, and indeed, we suspect that inadequate sampling limits accuracy for longer loops in many reported results. Here, we take a hierachical approach in which non-rate-limiting steps are allowed to scale exponentially, but the rate-limiting side-chain sampling/scoring stage scales linearly with loop length.

library. Instead, an effective sampling resolution is used in the following manner. For a single residue, the entire list of (ϕ,ψ) combinations is screened for steric clashes (and other screens described below). However, not every backbone conformation that survives the screening is retained for the buildup procedure. Rather, the screened rotamer states are further filtered, retaining a set of states in which all pairs of states obey the relation $\Delta\phi^2 + \Delta\psi^2 > R_{eff}^2$, where R_{eff} is the "effective resolution." This relation states that the distances, in the two-dimensional Ramachandran plot, between all retained pairs of backbone states must be larger than the chosen value of R_{eff}. Thus, this effective resolution allows pairs of retained backbone configurations to be identical in either phi or psi, but the combined changes in the two backbone angles must exceed the chosen threshold value. In our tests, this procedure holds a significant advantage over the conceptually much simpler procedure of creating, in advance, multiple dihedral angle libraries with varying resolutions. The major advantage occurs when a residue is located in a constrained position (i.e., steric restraints from surrounding residues), so that only a relatively small portion of the Ramachandran-allowed region is actually accessible. In such cases, the use of a coarse dihedral angle library could lead to zero acceptable backbone conformations being identified for the residue. The use of the very high resolution rotamer library permits all accessible regions of the dihedral angle space to be located, even very small ones, while the subsequent screening based on R_{eff} makes it possible to avoid retaining unnecessarily large numbers of conformations. Put simply, we construct an adaptive dihedral angle library for each residue "on the fly" based on its local environment.

The primary mechanism for screening loop conformations as they are being generated is identification of steric clashes. For computational efficiency, this screen is based on the ratio of distance between two atoms centers to the sum of their atomic radii [the "overlap factor" (*ofac*)], rather than directly computing the van der Waals interaction energy. The value of *ofac* is user-adjustable, but in the tests used here, we used both 0.75 and 0.7, and the results are then combined, as described in the section entitled "Hierarchical Refinement Strategy" below. In our experience, high-resolution crystal structures typically have *ofac* values no less than 0.75; physically, this value corresponds approximately to the point at which the Lennard–Jones 6–12 potential begins to increase rapidly. With respect to the algorithm, lower values of *ofac* tolerate larger steric clashes; these steric clashes can later be removed during the minimization stage, but lower values of *ofac* of course also cause a larger number of loop candidates to be generated. On the other hand, setting *ofac* too high can lead to the rejection of loop structures that are close to the native. This can be a particular problem when validating the algorithm against experimental structures that themselves contain moderate steric clashes (presumably due to incomplete model refinement). The values we have chosen work well for most purposes. The computational expense of searching for steric clashes is reduced through the use of a cell list.

Three other screens are performed in addition to identifying steric clashes. The first is designed to ensure that sufficient space exists for the side-chains on the loop. Otherwise, reasonable backbone configurations can result in one or more C^α–C^β vectors pointing in directions that do not permit the side-chains to fit sterically. This screen is accomplished by testing all conformations in a 30° side-chain rotamer library, until one is found that is free of steric clashes. The other side-chains of the loop are not included in these screens; only atoms on the body of the protein and the backbone of the loop are included. This screen does not guarantee that it is possible to pack the side-chains of the loop together in a combinatorial sense, but it does eliminate large numbers of incorrect loop conformations.

Loop conformations are also rejected as they are being constructed if the backbone of the loop travels too far away from the body of the protein. That is, most loops form contacts with the rest of the protein and only rarely project out into the solvent with no such contacts. We have quantified this phenomenon by evaluating the closest distances between loop backbone atoms and the remainder of the protein (including side-chains) in >500 proteins. Approximately 99% of the time, this minimum distance is <6.32 Å *for every* C^α *atom,* even for long loops, and we have used this cutoff value in all of the tests performed here (although it is user-adjustable). This criterion is applied to every residue in the loop. Residues at the two ends of the loop automatically satisfy this criterion because the adjacent residues (not on the loop) are within the cutoff distance. However, for residues in the middle of the loop, this criterion ensures that the loop cannot travel too far from the body of the protein.

Finally, loop conformations are also eliminated when it becomes clear that they will not be able to close. Once one or more residues have been built up from either side, an evaluation is made whether the distance between the end of this fragment and the other end of the loop is small enough that the remaining residues can geometrically span the gap. Here again, we have used a statistical analysis on loop regions in a test set of >500 proteins and determined the maximum distances spanned by various numbers of residues within the loop. At the 99.5% level, the maximum distance that can be spanned by four residues (C^α–C^α distance) is 13.97 Å, for example.

Loop Closure and Screening

Closed loops are generated by sampling conformations of the N- and C-terminal halves of the loop independently and then applying a loop closure algorithm in the middle of the loop. Currently, closure is accomplished as follows. The buildup procedure continues on either side of the loop up to the C^α atom on the closure residue. Then, all pairs of loop fragments built from the two sides are identified, which have the closure C^α atoms within 0.5 Å of each other (the cutoff value was determined empirically); that is, we identify all pairs of fragments that "meet in the middle." A

TABLE II. Critical Parameters for Loop Optimization, for Various Loop Sizes

Number of residues	4	5	6	7	8	9	10	11	12
Minimum loop conformations	16	32	64	128	256	512	1024	2048	4096
Minimum scored loops	16	20	24	28	32	36	40	44	48

closed loop is generated by averaging the positions of the closure C^{α} atom from the two fragments and adding the C^{β}, H^{α}, and side-chain atoms to the closure residue using standard geometries.

Not all closed loops are accepted to clustering. The closure procedure can generate loops with steric clashes and unacceptable geometries on the closure residue. The screens applied are the following:

1. N–C^{α}–C angle on the closure residue. The default cutoff value (from the "ideal" value of 111.1°) is 25°.
2. Backbone dihedral angles on the closure residue. Because the dihedral angles on the closure residue are not explicitly sampled and are determined instead by the geometries of the independently sampling fragments, we ensure that these dihedrals fall within the allowed portions of the Ramachandran plot. In practice, we allow the angles to fall slightly outside the allowed regions, by up to 25° (default value).
3. Steric clashes between the two halves of the loop. Because the halves are generated independently, there is no guarantee that they are compatible with each other.
4. Sufficient space for the side-chain on the closure residue. The rotamer library for the side-chain (see below) is scanned to ensure that at least one conformation of the side-chain lacks steric clashes with the body of the protein.

Closed loops that satisfy these screens are passed to the clustering algorithm.

Adaptive Control of Sampling Resolution

Although it is possible to choose the effective sampling resolution in advance, we have found that a simple adaptive method for choosing the sampling resolution for each loop individually is much more powerful. The critical issue is the wide variation in the "flexibility" of loops, even for constant loop length, which is correlated largely with the distance between the loop stems but also depends on other factors such as amino acid composition (especially number of Pro and Gly residues) and location within the protein. Using the same sampling resolution with floppy and tightly constrained loops can lead to enormously different numbers of generated loop conformations, particularly for long loops, which can create difficulties in terms of memory allocation, algorithmic speed, and effectiveness of the clustering algorithm (described below).

Our strategy is to define in advance the minimum and maximum number of loops to be generated for a particular loop length and then to gradually decrease the effective sampling resolution, beginning from a very coarse value, until the number of generated loops is intermediate between the minimum/maximum parameters. The maximum number of loops is fixed primarily on the basis of practical considerations, including memory and speed, and is set to 10^6 in this work. The minimum number is determined empirically by optimization on a diverse test set and varies with loop length, specifically according to 2^N, where N is the number of loop residues (Table II). This choice ensures a reasonable sampling of conformational space in all cases. It should be emphasized that the hierarchical refinement scheme ensures that low-energy basins on the energy surface are ultimately sampled much more finely than other portions of conformational space.

Constrained Optimization

The conformational space sampled by the loop buildup procedure can be constrained in two ways: restrictions on dihedral angles or restrictions on the Cartesian coordinates of the loop backbone atoms. We have found both types of constraints to be useful in the context of homology modeling. For example, long loops frequently have some small amounts of secondary structure embedded in them (e.g., a short helix or a few residues involved in strand pairing). These short secondary structure regions can in some cases be postulated to exist for sequences with unknown structure based on alignment to a template or based on secondary structure prediction algorithms. The loop prediction can then be constrained to maintain several residues in conformations that are consistent with the postulated secondary structure. In addition, in cases where the loop stems may be incorrect, the loop prediction can be extended into the adjacent secondary structure elements, but with dihedral angle or Cartesian constraints applied to decrease the computational expense of expanding the number of predicted residues and to ensure that the secondary structure assignment of the stem residues is not modified. The dihedral angle constraints have not been applied in our tests here, which focus on ab initio loop prediction, but they have been tested in the recent CASP5 competition.

One use of the Cartesian constraints is to prohibit the C^{α} atoms of the loop backbone from moving more than a prespecified distance from the initial loop conformation. This can be useful in the context of homology modeling when the loop conformation is believed to be well conserved between target and template based on functional information, multiple structure information for the relevant protein family, or local sequence conservation. In the tests performed here, the Cartesian constraints play a different role, as a mechanism for

focusing the sampling during the refinement phase of the loop prediction; the loop refinement stage is discussed in greater detail below.

Clustering

The dihedral angle-sampling buildup procedure can generate many thousands of loop candidate structures; the current implementation uses a maximum of 10^6 closed loops. These loops have already been screened to ensure that no steric clashes exist within the loop or between the loop and the body of the protein and that sufficient space exists for the side-chains. The side-chain positions have not been optimized, but in principle all of these loops could be successfully scored. By using the all-atom energy model it is possible to side-chain optimize and minimize hundreds or even thousands of loop candidates (with an average computational time of a few minutes per candidate for an eight-residue loop on a current generation PC). Even if it were possible to optimize and score each loop candidate, such a procedure would be computationally wasteful because many of the candidate loops would converge to similar final structures with similar energies.

Our solution is to use a clustering algorithm to select a representative set of loop candidates to optimize and score. The clustering algorithm we use is the K-means algorithm, as implemented by Hartigan and Wong.[29,30] A particularly salient feature of this algorithm, for our purposes, is that the computational expense 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. (The algorithm attempts to minimize the within-cluster sum-of-squares, but it cannot guarantee a globally optimal solution. In practice, the locally optimal solutions are entirely adequate for our purposes.) 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 4 times the number of loop residues (Table II).* The descriptors used for clustering are the Cartesian coordinates of the N, C^{α}, C^{β}, and C atoms for each loop residue. Used in this way, the algorithm rarely requires more than a few seconds of computational time, even with tens or hundreds of thousands of loop structures. The initial cluster centers used to initiate the algorithm are generated as suggested by Hartigan and Wong.[30]

We have also implemented a simple method of adaptively increasing the number of clusters when it appears to be warranted. Specifically, if after initial clustering the variance of one or more clusters is much larger than the others (specifically, more than 4 times the variance of the cluster with the median variance), then these clusters are "split" into two new clusters, and the clustering algorithm is run again with the new clusters added. This procedure can be applied iteratively, but in the results presented here, at most, one iteration of "cluster splitting" was applied. In the vast majority of cases, this simple procedure succeeds at greatly reducing the spread of the cluster variances (i.e., balancing the size of the different clusters and giving a more uniform cluster distribution).

Cluster representatives (i.e., representative loops from within a particular cluster that are chosen for optimization) are chosen according to their distance from the cluster center, with more central representatives chosen first.

Side-Chain Optimization

Loop structures chosen for scoring are first subjected to side-chain optimization. Our algorithm for side-chain optimization has been described previously[1] and is summarized here. Sampling is accomplished primarily by using a highly detailed (10° resolution) rotamer library constructed by Xiang and Honig[31] from a database of 297 proteins. This library contains, for example, 2086 rotamers for lysine. The computational expense of such a detailed library was mitigated by prescreening the rotamers by using only hard sphere overlap as a criterion (using a cell list for computational efficiency), allowing many rotamers to be excluded before performing any energy evaluations. The method we use for the combinatorial optimization is also adapted from the method of Xiang and Honig,[31] which is similar in spirit to earlier work by Bruccoleri and Karplus.[14] 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, all side-chains are completely energy-minimized simultaneously in Cartesian coordinates to remove any remaining clashes.

Minimization

After side-chain optimization, the loop, including all side-chains, is completely energy-minimized ($<$0.001 kcal/mol/Å final root-mean-square gradient) in Cartesian space (i.e., all atoms are free to move) using a novel multiscale minimization algorithm (M.P.J. and R.A.F., unpublished results). This algorithm is a variant of the truncated Newton (TN) method, specifically the TNPACK implementation of Schlick and coworkers.[32–34] The implementation is based on a division of the molecular mechanics forces into short- and long-range components, in analogy to multiscale molecular dynamics methods such as RESPA.

*The number of clusters, which is equal to the number of loops that are ultimately ranked using the all-atom energy function (after side-chain optimization and minimization), may seem small. We believe there are several reasons for the success of our chosen linear scaling of the number of clusters with loop length. First, the clustering algorithm generally performs extremely well in retaining low RMSD loops even when the number of clusters is substantially less than the number of conformations. Second, the hierarchical refinement procedure allows the algorithm to converge to low RMSD ($<$1 Å) as long as at least one loop conformation with reasonable RMSD (2–3 Å) is scored during the initial loop buildup, in our experience. The hierarchical refinement in turn relies on correlation between energy and RMSD, which is observed in most cases (exemplified by 11kk in Fig. 1), but only after extensive side-chain optimization and energy minimization.

Short-range forces include all bond, angle, and torsion terms in the force field, as well as all nonbonded interactions between atoms separated by <10 Å. The remaining nonbonded interactions constitute the long-range forces. The long-range forces are never evaluated during the "inner" TN cycles (which determine the line search direction) or during the line search and are only periodically updated in the outer TN cycles (in this work, once every five Newton cycles). The division of the nonbonded interactions into short- and long-range is also updated every five Newton cycles.

The Generalized Born (GB) solvation model is well suited for performing rapid minimizations because the pair screening term is analytical and thus differentiable. However, the resultant expression for the gradient involves derivatives of the Born α-values with respect to the atomic coordinates, which must be determined numerically. Because these derivatives are expensive to compute, we use a self-consistent minimization, in which the Born α-values are held fixed during the course of the minimization and then updated prior to another minimization, and so on until the energy ceases to decrease by >1 kcal/mol. In practice, self-consistency rarely requires more than two cycles of TN minimization, and the second minimization is generally extremely rapid (i.e., only a very small number of Newton cycles, with the energy typically changing by only 0.01–0.1 kcal/mol). This self-consistent minimization with GB solvent requires only ~50% greater computational expense than vacuum minimizations.

Simulation of Crystal-Packing Environment

When desired, crystal unit cells are explicitly reconstructed by using the dimensions and space group reported in the Protein Data Bank files. For most proteins, the crystal unit cell contains too many atoms for explicit lattice summation techniques (e.g., Ewald summation) to be computationally feasible. Instead, the simulation system consists of one asymmetric unit (which may contain more than one protein chain) and all atoms from other, surrounding asymmetric units that are within 20 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. A new version of the SGB/NP solvation free energy code was written to account for the crystallographic symmetry.

Details of Energy Function

The all-atom OPLS force field[25,26] is used to describe the protein intramolecular energetics. All nonbonded interactions between pairs of atoms within 20 Å of each other were calculated. The OPLS torsional energy parameters were recently refined by using high-level quantum chemical calculations[26] and validated by using protein side-chain prediction[24]; the updated parameters are used here. The solvation free energy was estimated by using an implicit solvent model consisting of the Surface Generalized Born (SGB) model of polar solvation[28] and a nonpolar estimator developed by Gallicchio and coworkers.[27] Correction terms have also been developed to improve the agreement between SGB and Poisson–Boltzmann solvation free energy calculations.[28]

Two surfaces (extending over the entire simulation region, including the symmetry copies) were calculated to perform the SGB surface integrals: one high resolution (330 points per sphere) and one low resolution (10 points per sphere). The distributions of points on the spheres used to construct these surfaces were determined by using the spiral points algorithm.[35] Because the magnitude of the integrand of the surface integral decreases rapidly with distance, the integration was performed with the high-resolution surface for all points within 7.5 Å of the charge in question, and the lower resolution surface at longer distances, up to an absolute cutoff of 20 Å.

Hierarchical Refinement Strategy

In the work reported here, the loop prediction algorithm described above was applied multiple times to each loop. The goals were twofold: 1) reduce the sensitivity of the results to the value of the "overlap factor" used to define steric clashes and 2) improve the sampling in low-energy basins in a hierarchical, iterative fashion. With respect to the former, two initial loop prediction runs are used with overlap factors of 0.7 (less stringent) and 0.75 (more stringent). With respect to the latter, refinement of the initial results proceeds in two stages. First, the best (i.e., lowest energy) five loops from each initial loop prediction are subjected to further sampling using a Cartesian restraint of 4 Å on each backbone C^{α} atom. Then, the results of all the initial and refinement predictions are combined, and the five lowest energy loops are subjected to a final stage of refinement by using tighter Cartesian restraints of only 2 Å. The refinement stages are, of course, optional, and in practice only lead to substantial increases in accuracy for loops of length eight residues or longer.

Data Sets and Criteria for Evaluation of Accuracy

We have investigated loops with lengths between 4 and 12 residues. To facilitate comparison with previous work, we have chosen to study loop databases compiled previously by Fiser et al.[9] and Xiang et al.[8]. The loops were chosen in these previous studies from databases of high quality (2.0 Å resolution or better), diverse (maximum sequence identity 60% for Fiser et al. and 20% for Xiang et al.) crystal structures. Fiser et al. chose 40 loops of each length studied, with no more than one from each PDB structure. Xiang et al. investigated all loops within their chosen protein data set; thus, the number of loops varies inversely with loop length due to natural abundance. Our results are reported as "global" RMSDs, obtained by superimposing the body of the protein; this choice is the same as that adopted by Xiang et al.[8] and de Bakker et al.[23] but differs from the "local" RMSD measure used by Fiser et al.,[9] which provides no information about the orientation of the loop with respect to the body of the protein. If the objective is to use the structures for understanding interactions relevant to biological function and medicinal chemistry (e.g., docking candidate inhibitors),

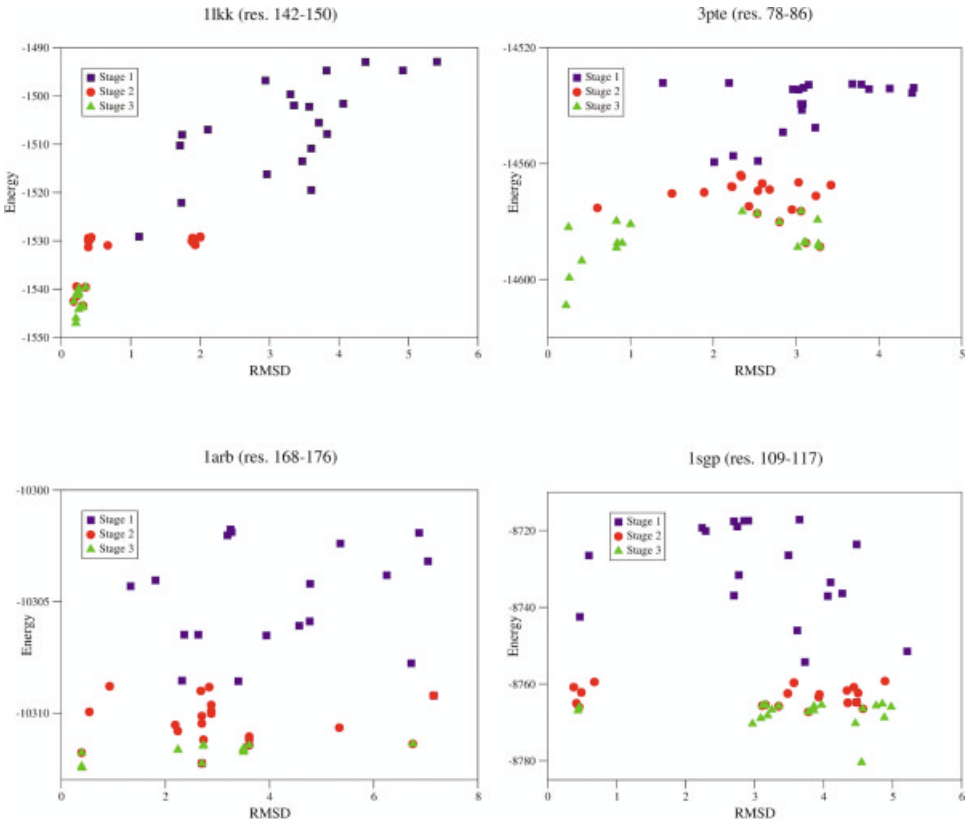

Fig. 1. Plots of energy (kcal/mol) versus backbone RMSD for local minima identified during the optimization of four nine-residue loops. The local minima are colored according to the stage during which they were found: stage 1 (initial optimization) is represented by blue squares, stage 2 (first refinement stage) by red circles, and stage 3 (final refinement) by green triangles. Only the 20 lowest energy minima from each stage are included in these plots.

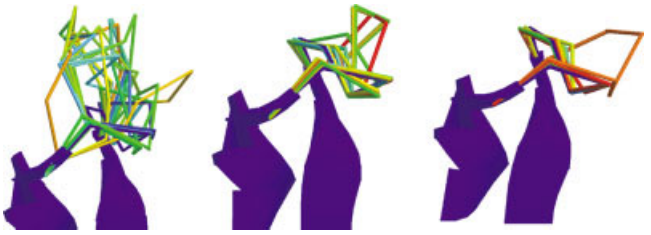

Fig. 2. Twenty lowest energy loops for 1lkk (residues 142–150) during stage 1 (left), stage 2 (middle), and stage 3 (right) of the optimization procedure. The energies of these loops are plotted in Figure 1.

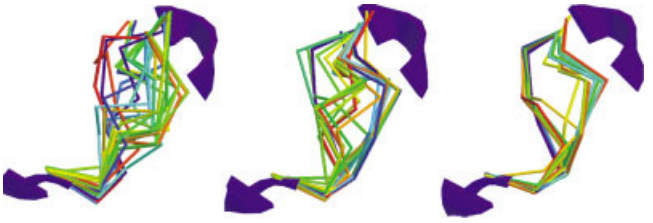

Fig. 3. Twenty lowest energy loops for 3pte (residues 78–86) during stage 1 (left), stage 2 (middle), and stage 3 (right) of the optimization procedure. The energies of these loops are plotted in Figure 1.

then accurate placement of the atoms of the loop in the context of the protein is of critical importance, and RMSD computed via loop stem superposition provides a reasonable measure of the effectiveness of the algorithm in accomplishing this task (ultimately, more direct measures, such as accuracy of docking into the resulting structure, would be even more relevant and insightful).

All of the loops in the Xiang et al. test set (ranging between 5 and 12 residues) are considered in this article. However, those of Fiser et al. are investigated only up to and including 10 residues in length. We have found that the loops in this test set with 11 or 12 residues in general contain a high percentage of secondary structure—so high that many of the test cases involve prediction of, for example, major components of β-sheets. This qualitative change in the composition of the data set is reflected in the very large (~6 Å) RMSDs reported by Fiser et al. in comparing their predictions of these test cases to experiment. Although such predictions are a potentially important aspect of the general problem of ab initio structure prediction, they become much more algorithmically tractable if specialized sampling algorithms are used (e.g., to systematically sample candidate regions with regard to strand formation and pairing, a relatively inexpensive effort as the phase space is drastically restricted). We have not implemented algorithms of this type in our present methodology. Furthermore, prediction of major architectural core elements of the protein structure is not typically

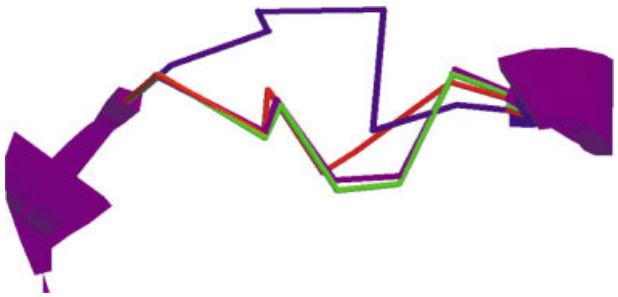

Fig. 4. Sequential refinement leading to the lowest energy loop in 3pte (residues 78–86). The green backbone trace is the lowest energy loop at stage 3 (final refinement), and the native structure is in purple (nearly superimposed). The lowest energy loop was generated by restricted sampling performed on the red loop (stage 2), which in turn resulted from restricted sampling performed on the blue loop (from stage 1, the initial loop generation stage).

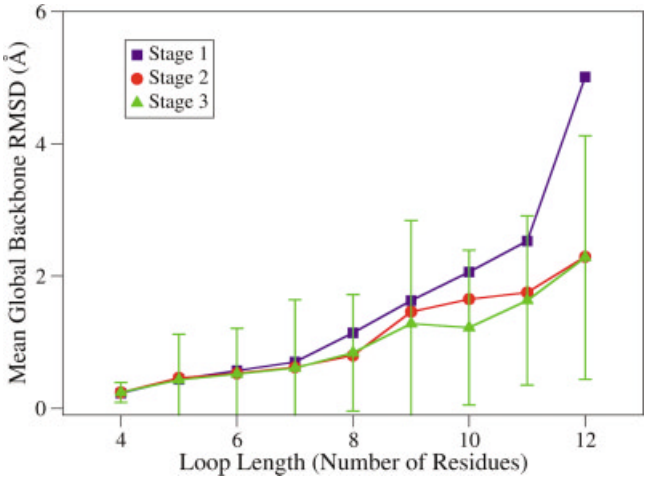

Fig. 5. Mean backbone RMSD for the combined filtered data set as a function of loop length and stage of the prediction algorithm. For the final (stage 3) results, the vertical bars indicate 1 SD.

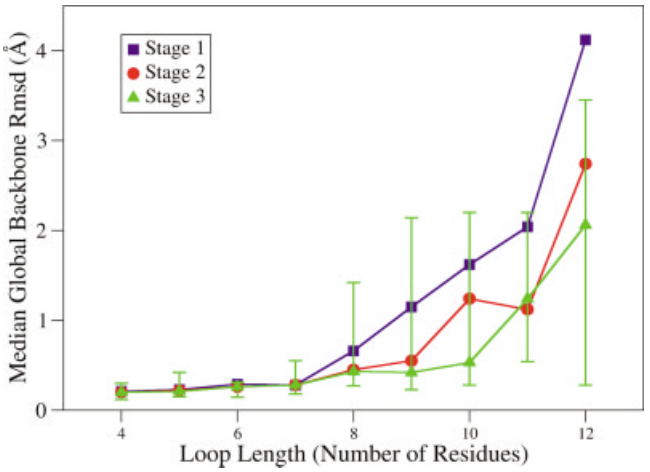

Fig. 6. Median backbone RMSD for the combined/filtered data set as a function of loop length and stage of the prediction algorithm. For the final (stage 3) results, the vertical bars indicate results at the 20th and 80th percentiles.

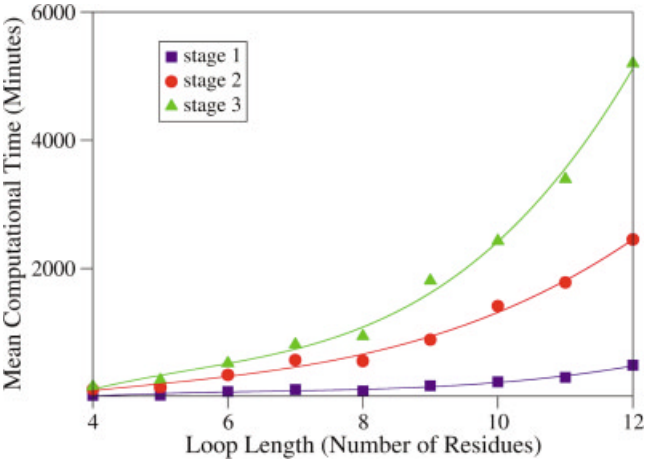

Fig. 7. Scaling of computational expense with loop length and as a function of the number of prediction stages performed. The lines represent a fit to a cubic polynomial.

required in many homology-modeling applications. Among the remaining loops in the data set, some have a few residues in a β-strand or α-helical configuration, typically at the ends of the loops (and occasionally in the middle as well); however, the reconstruction of a small amount of secondary structure in such locations is common in realistic homology-modeling scenarios, and we retain these cases. The one exception is residues 8–18 in the protein 1msi in the 11-residue data set of Xiang et al., which contains a three-residue β-strand in the middle of the loop that is a part of a three-stranded sheet. This case is included in our test set but is not included in the "filtered" database (see below for an extensive discussion of other filtering criteria) for this architectural reason.

RESULTS AND DISCUSSION

To illustrate how the hierarchical loop prediction algorithm works in practice, we first analyze in detail the results for a few loops before discussing overall results on a large data set.

Detailed Analysis of Selected Illustrative Cases

Figure 1 illustrates the hierarchical loop prediction procedure for four nine-residue loops. The 1lkk case is typical of most loops of similar length. A strong correlation can be observed between energy and RMSD. The initial loop buildup algorithm generates several loops within 2 Å RMSD of the native, and the subsequent refinement stages succeed in generating several loops within 0.5 Å RMSD of the native, and these in fact have the lowest energies. In this particular case, the final refinement stage succeeds in lowering the energy further, but not the RMSD. The loop structures associated with the points in the energy versus RMSD plot are represented in Figure 2. In this well-behaved case, the conformational space examined at each stage of the optimization narrows in a simple manner around the native loop.

The 3pte case is similar, except that there is clear evidence of two basins of attraction: one corresponding to

the native and the other near 3 Å RMSD. The local minima identified during "stage 1" show no strong correlation between energy and RMSD; the correlation becomes evident only when the local minima from the refinement stages are considered. Here, the final refinement (stage 3) is critical to the attainment of high accuracy. The lowest energy minima after stage 2 correspond to the competing, incorrect basin of attraction. However, as can be seen in Figure 3, the second stage of refinement clearly samples both basins of attraction and establishes the native basin of attraction as lower in energy. Finally, Figure 4 presents the hierarchical refinement for this loop in a different manner, by tracing the "history" of the lowest energy loop generated. That is, the green backbone trace is the lowest energy loop at stage 3 (final refinement), and the native structure is in purple (nearly superimposed). The lowest energy loop was generated by restricted sampling performed on the red loop (stage 2), which in turn resulted from restricted sampling performed on the blue loop (from stage 1, the initial loop generation stage). Further details of the sampling for this case are also presented in Table I, highlighting the individual steps in the algorithm: initial dihedral angle buildup, screening, clustering, and side-chain optimization/energy minimization.

The 1arb case is also challenging and represents a class of loops (relatively small in number) in which multiple low energy, nearly degenerate basins of attraction can be identified. The lack of clear correlation between energy and RMSD suggests that this loop may be relatively "floppy," in the sense that relatively minor perturbations to the environment may allow it to adopt alternate conformations. Here again, the refinement stages are critical to obtaining a low RMSD result.

Finally, the 1sgp case is representative of those cases (rare among the nine-residue loops) where the sampling algorithm succeeds in generating low RMSD structures, but the energy function clearly favors a competing, incorrect basin of attraction. These cases of course do not automatically suggest that the energy function is inadequate. No ligands lie close to this particular loop, but errors in protonation states are certainly possible, either on the loop itself or nearby.

Factors Affecting Accuracy of Predictions

The protocol we use here to evaluate accuracy of loop prediction mandates reconstruction of a loop of specified length into the native protein structure. Success in this endeavor is a necessary condition for an accurate structural refinement methodology; however, it is not sufficient to guarantee good results in a more realistic homology-modeling context, in which the side-chains proximate to the loop would not necessarily be accurately positioned. Furthermore, in many realistic homology-modeling scenarios, the alignment used to generate the model cannot be assumed to be perfect, and resultant errors in the remainder of the structure can complicate loop prediction. The amount of degradation introduced into the results by distortions of the remainder of the protein from the native structure is not at this point clear, in our work or in the work of other groups (although we have obtained some limited preliminary results suggesting that prediction of neighboring side-chains leads to a relatively modest increase in error, on the order of 0.1 Å, for eight-residue loops). This issue will be the subject of extensive investigations in the future.

We present results for all remaining loops, 833 in total. However, in an effort to assess the origin of errors in the results, we also consider the following three criteria:

1. The pH at which the protein was crystallized
2. Whether the loop interacts with heteroatom species, principally ligands or ions, that are a part of the X-ray crystal structure
3. The quality of the crystal structure in the prediction region.

The criteria for segregating test cases from the data set due to each of these three factors (all of which are automated and based on physically reasonable arguments, to avoid bias) are discussed in detail below. A resulting data set, with these cases removed, is referred to as the "filtered" data set, and we view the associated statistics as the most accurate representation of the current performance of the methodology. Nonetheless, we also report for comparison results for the entire ("unfiltered") data set; individual results for each protein, whether "filtered" or not, are available in the Supplementary Material.

The pH has a strong effect on the protonation states of ionizable groups; however, in the present work, we have chosen (for simplicity) to use only standard protonation states of every residue (Asp, Glu ionized; Arg, Lys protonated; His neutral). One can of course on occasion observe alternative protonation states at neutral pH (particularly of His) and this can have a significant effect on the accuracy of loop prediction. However, prediction of alternative protonation states at neutral pH would require enumeration of the various possibilities, simulations with each protonation state, and final assessment of the relative free energies of the various candidate loops generated with the different protonations (amounting in essence to a methodology for pKa prediction as well as loop prediction). This challenge will be addressed in future work but not in the results presented here. We have chosen in this article to segregate loops that 1) belong to PDB structures reported to have been crystallized at a significantly nonstandard pH and 2) contain an ionizable group (or interact with such a group) that would be affected by the change in pH (typically, carboxylates at low pH which are then easily protonated).

In principle, interactions with ligands or ions can be modeled in the same fashion as one models the interactions of various residues of the protein. In practice, we have not optimized potential function (or solvation) parameters for general ligands or ions, and our intention in this initial article is to focus on the performance of the methodology and energy function in a well-defined regime where the model has already been subjected to extensive testing and optimization (via, e.g., single side-chain prediction)

and have simply not included ligands or ions in our calculations. In cases where the interactions with these groups are extensive, large errors can obviously be expected when the relevant atoms are not incorporated at all into the model.

We have assessed the issue of whether any given prediction region is substantially affected by interaction with a ligand or ion in the following automated fashion. The closest (heavy atom) distance is calculated between any atom in the ligand and any atom in the protein. A neutral ligand is considered to closely interact with the loop if the distance of closest approach is 4.0 Å or less (i.e., van der Waals contact for heavy atoms in the two first rows of the periodic table). Organic ions are treated in the same fashion. For metal ions, effects can extend beyond side-chains ligated to the metal into the second coordination shell. The reason is that a typical constitutive ion such as Zn or Cu will have a number of charged side-chain ligands. If the metal is removed and has a net charge (as is ordinarily the case), then the charge on the side-chains will (incorrectly) no longer be neutralized by the metal, and the resultant electrostatic effects can easily extend out into the second coordination shell. Therefore, we set the interaction cutoff for metal ions at 6.5 Å. Predictions for regions interacting with ligands and/or ions will be considered in greater detail in future work, when the required parameters have been incorporated into the loop prediction code and the charge states of the metal ions have been properly analyzed.

Assessing the quality of experimental structures is difficult, and we do not intend to undertake an exhaustive analysis. However, one aspect of structural quality that can have a very direct affect on loop reconstruction results is the presence of heavy atom overlaps involving the loop in the PDB file. These overlaps make it difficult to redock the targeted prediction region back into the native protein without relaxing that structure, something that is not a part of the protocol used to generate the results in this article. If we were selecting structures from scratch to build up our own data set, we would simply set a cutoff for the maximal degree of overlap and reject any loops with greater overlap than is specified by the cutoff. In the present work, we segregate cases in which the minimal "overlap factor" for any heavy atom in the loop is <0.7, which is the same cutoff that we use in our loop generation algorithm. The overlap factor, which is used extensively as a filter during the loop buildup procedure, is discussed in detail in Materials and Methods. Briefly, it is defined as the ratio of the distance between two atomic centers to the sum of their radii; an overlap factor of ≤ 0.7 represents a very serious steric clash. Only about 1–2% of the loops have such a serious steric clash, and a disproportionate number of these have exceptionally high RMSDs, due in many cases to sampling errors; that is, the "native" structure cannot be reconstructed because our algorithm prohibits the creation of such steric clashes. Steric clashes in the "native" loop can also lead to energy errors in that, if the packing of the loop is too tight from the point of view of the potential energy function, it may be able to squeeze the loop into the required space, but at the cost of incurring an energetic penalty via repulsive van der Waals interactions.

TABLE III. Number of Loops for the Combined Full and Filtered Data Sets

	4	5	6	7	8	9	10	11	12
Full	40	160	147	114	101	97	71	37	21
Filtered	35	117	100	82	66	57	40	18	10

Finally, highly disordered loops are identified by using the average "B-factor" for the backbone atoms in the loop ($N,C,C^{\alpha},C^{\beta},O$). We choose a cutoff of 35; although any such cutoff is somewhat arbitrary, very few loops in our test set fall into this category, and as might be expected, their prediction accuracy is significantly degraded from loops with less disorder.

In reporting results for individual loops (Supplementary Materials), we indicate whether each loop possesses one or more of the above features, quantifying and characterizing the feature (e.g., identifying a perturbing ligand or ion and reporting the distance of closest approach of the prediction region to a heteroatom) as appropriate. We then reported statistics for the entire data set (unfiltered) and for loops that pass the various filtering protocols (the filtered data set). The following statistics are reported for each loop length

1. Number of example loops (Table III) for the combined Fiser et al. and Xiang et al. data set. We focus our discussion on this combined data set in what follows, because the principal objective of the article is comparison with experiment rather than with prior work. However, results broken out separately for the individual data sets are provided in the supplementary material.
2. Average and median RMSDs are reported for the final loop prediction for both the full and filtered data sets (Table IV). Results are also reported at the three iterative stages of the loop prediction algorithm for the filtered dataset (Figs. 5 and 6). Reporting results at each stage of the algorithm allows the effectiveness of the iterative protocol to be assessed, an evaluation that is meaningful only for the filtered data set (there is no reason to expect loops with, e.g., strong interactions with missing ligands or ions to improve with iteration).
3. Fraction of cases exhibiting energy errors (Table V) (lowest predicted loop energy less than the optimized native loop energy) and sampling errors (Table VI) (lowest predicted loop energy greater than the optimized native loop energy) as a function of iteration for loops with an RMSD > 1.0 Å above the median RMSD. These values indicate the percentage of outliers in the data set that are due to the model as opposed to quality of coverage of phase space. We also present the average value of the energy gap for structural outliers for both sampling and energy errors (Table VII). Again, only the filtered data set is considered for reasons analogous to those indicated above.

TABLE IV. Mean and Median Global Backbone RMSDs for the Combined Data Set

	4	5	6	7	8	9	10	11	12
Full/mean	0.24	0.44	0.59	0.77	0.98	1.37	1.70	2.66	2.67
Full/median	0.20	0.24	0.28	0.30	0.44	0.51	1.09	1.87	1.93
Filtered/mean	0.24	0.43	0.52	0.61	0.84	1.28	1.22	1.63	2.28
Filtered/median	0.20	0.21	0.26	0.28	0.43	0.42	0.53	1.24	2.06

TABLE V. Fraction of Errors Attributable to the Energy Function for the Combined/Filtered Data Set

	4	5	6	7	8	9	10	11	12
Stage 1	0.00	0.05	0.07	0.11	0.12	0.14	0.23	0.17	0.00
Stage 2	0.00	0.07	0.10	0.10	0.15	0.25	0.30	0.33	0.10
Stage 3	0.00	0.08	0.10	0.11	0.17	0.28	0.33	0.39	0.20

TABLE VI. Fraction of Errors Attributable to Sampling for the Combined/Filtered Data Set

	4	5	6	7	8	9	10	11	12
Stage 1	0.00	0.03	0.05	0.06	0.26	0.40	0.38	0.72	0.90
Stage 2	0.00	0.03	0.03	0.02	0.09	0.16	0.23	0.22	0.60
Stage 3	0.00	0.02	0.02	0.01	0.11	0.09	0.10	0.22	0.40

TABLE VII. Average Size of Sampling and Energy Errors and Their Maximum and Minimum Values (for the Combined/Filtered Data Set)

	4	5	6	7	8	9	10	11	12
$<+\Delta E>$	N/A	10.85	3.65	7.90	3.96	19.22	14.78	8.48	23.20
Max + ΔE	N/A	16.20	4.90	7.90	11.90	45.90	35.10	14.40	39.60
$<-\Delta E>$	N/A	−9.20	−15.36	−12.63	−19.85	−13.70	−20.69	−15.99	−18.40
Min $-\Delta E$	N/A	−22.90	−41.10	−32.30	−67.10	−35.70	−34.40	−43.20	−27.70

4. Average CPU times as a function of iteration stage and loop length (Fig. 7). This is presented for the full data set, because the problems mentioned above have minimal impact on computational effort.

Accuracy of Loop Prediction

As mentioned above, the discussion that follows focuses on the filtered data set, because that is what reflects (as argued above) the actual characteristics of our energy function and sampling algorithm, as opposed to qualitatively missing or incorrect features of the physical model. The unfiltered results display improvement (in some cases considerable improvement) compared with those reported by the original authors. However, even these comparisons are nontrivial to evaluate with regard to model accuracy and/or sampling effectiveness, because we include crystal-packing effects, which have been ignored by previous studies. Hence, the main focus of the analysis is on assessing the level of prediction accuracy that has been reached in the filtered data set in an absolute sense (i.e., compared with experiment).

An in-depth examination of the results suggests that characteristics of the loop prediction problem are not uniform as one goes from relatively short loops (5–8) to relatively long loops (11–12). Our predictions for short loops are in general remarkably accurate and robust, with very few outliers, especially within the "filtered" data sets. This is manifested quantitatively in Table I via the factor of >2 disparity between the median and average RMSD (a disparity in fact persists through loops of length 10). Both the average and median RMSDs increase more or less linearly as a function of loop length; the median RMSD, in the range of 0.2–0.4 Å, is arguably not very far from the noise of the experimental structures (which, it should be remembered, themselves require modeling to produce).

For short loops, the algorithm displays very few large sampling errors (i.e., cases where the energy gap between the predicted structure and the optimized native structure is large and positive), and the small number of outliers in RMSD are predominantly due to energy errors, some of which can be quite large (although generally not as large as for the longer loops). There are a number of possible sources of these errors:

1. Nonstandard protonation states in the loop or surrounding side-chains of the protein, in cases where there is no indication of nonstandard pH in the PDB file. A good example of this type is 1brt in the seven-residue Xiang et al. data set, which has an RMSD of 6.73 Å from the native and an energy error of 23.1 kcal/mol. In the native structure of this loop, there is a salt bridge formed between an aspartic acid and histidine residue, implying that one or other of the side-chains must be protonated, despite the fact that the pH is reported to

be 8.5. Without protonation, there is no way to form the salt bridge, and this in turn may drastically alter the lowest energy loop conformation.
2. Structures sampling unusual regions of the potential energy surface, which may be qualitatively inaccurate (clearly, a high percentage of that surface must be reasonably accurate; otherwise, one would not see such good agreement in both RMSD and energy of the predicted and native structures for a substantial majority of test cases);
3. Systematic errors in the solvation model leading to a preference of one type of structure (e.g., solvent exposed) over another.

Analysis of the energy errors in the outliers will enable improvement of the potential function and solvation model.

Loops of length 11–12, on the other hand, appear to be qualitatively more challenging with regard to prediction than those in the 5- to 8-residue category. Some of this is due to the greater difficulty of the sampling problem, which is manifested in Table VI by the increase in the fraction of sampling errors for outlier loops as loop length increases. On top of the basic exponential combinatorial explosion in the number of possible loop conformations (which our sampling algorithm ameliorates but does not entirely neutralize), long loops tend to lie farther from the body of the protein, increasing the number of alternative conformations that avoid steric clashes. It is also the case that the absolute number of outlier energy errors increases with loop size, as well as the average size of the energy gap. A possible explanation for this phenomenon is that, as the loops become longer and acquire more "play," they have increased opportunity to locate regions of the energy function where there are substantial errors; again, these structures can be used to track down problems and ultimately improve the molecular mechanics potential and solvation model. Finally, we note that a relatively large number of loops of these lengths in the Xiang et al. data sets have heteroatom interactions, so that the filtered data sets are considerably smaller than is optimal for statistical analysis, particularly for the 12-residue loops where the data must at present be considered anecdotal. Thus, the conclusions for these longer loops should be regarded as provisional, and studies on larger data sets will have to be conducted to achieve a reliable estimate of accuracy at these lengths. If the 12-residue loop results are indeed representative, we observe that at this length the accuracy of the methodology becomes problematic in general (as opposed to for a small number of outliers), suggesting that significant improvements in both sampling methodology and the energy model will be needed to produce robust results for loops of this size and larger.

Loops of length 9–10 display an intermediate behavior. There is a jump in the average RMSD, a few examples of significant sampling failure and energy failure (1wer, where the energy of the predicted structure is 32 kcal/mol below that of the native structure), and generally more loops that experience difficulties of one sort or another. Nevertheless, the performance overall is highly encouraging and substantially better than previous results. Here, the size of the combined Fiser et al./Xiang et al. data set is sufficiently large that the filtered data sets are of adequate size, so the conclusion that the methodology is generally quite accurate for these cases, with an occasional serious problem, is well established.

The results in Figures 5 and 6 and Tables 8 and 9 show that, for the longer loops, the iterative prediction methodology is essential to achieve good results. Starting with loops as short as nine residues, improvements of 60% or more are obtained going from the first to the third iteration. Indeed, for nine residue loops, there are very few clear-cut sampling errors remaining by the third iteration (as opposed to the first, where there are quite a few), and those that are manifest are in some cases due to failure to form optimal β-strand pairs or sheets. The iterative algorithm can be modified in its details (e.g., values of the Cartesian restraints, number of iterative cycles), and effort along these lines may be needed to improve results for longer loops. Finally, modification of the loop prediction algorithm to specifically explore the strand-pairing region of phase space is straightforward and will be reported in a subsequent publication.

As was mentioned above, it is difficult to make rigorous comparisons with the previous studies from which our data set is drawn[8,9] because of our inclusion of crystal packing and the issues involving filtering of the data sets. However, a number of qualitative observations can be made. First, it is clear that the results of Fiser et al.,[9] in which a statistical database potential is used, are qualitatively inferior to the present results, for the loop lengths considered here, 4–12 residues. This finding suggests that the use of statistical database potentials, as opposed to realistic models based on physical chemical principles, has a fundamental limitation in achieving high-resolution structural refinement. Other published results using such potentials to carry out loop predictions led to similar conclusions; in particular, the recent results of de Bakker et al.[23] have also suggested that all-atom force fields are capable of qualitative improvements in accuracy relative to statistical potentials. It is difficult to see how the fine details of flexible loop structure can be captured by a potential function that in essence ignores solvation effects and electrostatics and incorporates conformational analysis only at a very approximate level.

The energy function of Xiang et al.[8] can be viewed as one that heuristically models solvation and electrostatic effects; as such, it is intermediate between the all-atom energy function used here and statistical energy functions, in both speed and accuracy. The greater speed achieved by heuristic models (roughly 1–2 orders of magnitude compared to the CPU times we report below) can be important in, for example, rapidly providing a good initial guess for a homology model, which could then be refined by more accurate methods. Xiang et al.[8] also introduced a novel means of approximately introducing effects of entropy without extensive free energy sampling: the so-called "colony energy," which uses clustering of the structures generated by the sampling algorithm as a constituent of

TABLE VIII. Summary Table of Results for the Fiser et al.[9] Nine-Residue Loop Test Set.

				Stage 1				Stage 2				Stage 3						
#	PDB	RESLO	RESHI	BACK	HEAVY	dE	TIME	BACK	HEAVY	dE	TIME	BACK	HEAVY	dE	TIME	B	HET	OFAC
1	1npk	_:102	_:110	0.1	0.5	−1	2.3	0.1	0.5	−1	7.5	0.1	0.5	−1	29.7	14		
2	1arp	_:127	_:135	0.3	1.0	8	1.5	0.1	0.8	−7	8.5	0.1	0.8	−7	29.0	13	7.4	
3	2fox	_:5	_:13	3.0	3.0	6	2.0	1.5	1.8	1	3.8	0.2	0.8	−1	20.4	10	2.5	
4	2alp	_:139	_:159	0.7	2.0	35	1.1	0.3	1.0	17	6.6	0.2	0.7	12	7.1	8	6.5	0.75
5	1onc	_:70	_:78	1.4	2.3	2	3.3	0.2	2.3	−20	20.4	0.2	2.3	−20	20.4	13	13.5	
6	2cpl	_:24	_:32	0.1	0.9	9	2.0	0.2	0.7	−4	10.4	0.2	0.7	−4	17.8	27		
7	1cyo	_:49	_:57	7.5	8.8	49	1.1	0.6	1.3	19	5.7	0.2	1.0	0	7.9	11	3.6	
8	4gcr	_:94	_:102	0.4	1.9	17	2.7	0.5	1.8	2	2.7	0.2	0.9	−4	80.4		3	
9	1php	_:91	_:99	3.8	4.5	18	2.3	0.4	1.3	−1	11.2	0.3	1.1	−5	27.9	17	29.7	
10	1mrk	_:53	_:61	0.3	0.8	14	1.9	0.2	0.5	1	10.0	0.3	1.1	0	10.0	19	11.8	
11	2sil	_:183	_:191	1.3	2.3	25	2.6	0.3	1.3	6	11.5	0.3	1.2	0	42.6	13		
12	5fx2	_:8	_:16	0.0	0.0	25	1.5	0.0	0.0	1	1.5	0.3	0.9	1	7.7	10	2.1	0.68
13	2dri	_:130	_:138	0.3	0.6	−1	3.5	0.3	0.6	−3	9.4	0.3	0.6	−3	26.6	14	3.2	
14	1noa	_:76	_:84	3.1	4.9	21	1.2	0.3	1.2	1	7.3	0.4	1.4	−2	7.3	18		
15	1btl	_:102	_:110	1.5	4.1	6	2.5	0.7	1.2	−7	9.9	0.4	1.1	−14	26.7	14	5	
16	1csh	_:252	_:260	0.4	0.8	−5	8.5	0.4	0.6	−6	121.8	0.4	0.6	−6	162.0	8	16.8	
17	1xnb	_:116	_:124	0.4	1.5	−3	1.9	0.4	0.7	−9	10.6	0.4	0.7	−11	32.0	16	22.8	
18	1pgs	_:117	_:125	0.8	0.9	9	1.6	0.5	1.5	7	11.1	0.4	1.3	3	11.1	19		
19	1pda	_:108	_:116	0.5	1.1	−13	2.8	0.6	1.5	−17	9.9	0.5	1.2	−17	17.0	27	8.2	
20	1fnd	_:121	_:129	2.6	3.4	11	1.6	2.7	2.9	2	7.2	0.6	1.4	−14	10.0	26	4.5	0.7
21	1lif	_:73	_:81	2.4	3.4	100	4.9	0.5	0.8	100	21.2	0.6	0.7	100	33.8	14	3.6	
22	1aba	_:69	_:77	1.6	2.0	8	1.1	0.9	1.6	5	5.9	0.6	1.0	−1	16.2	14	11	
23	1amp	_:57	_:65	0.7	1.0	−1	3.5	0.6	0.9	−7	13.8	0.7	1.1	−7	16.6	30	21	0.72
24	1flp	_:41	_:49	3.6	4.6	19	1.5	1.1	1.5	2	6.9	0.8	1.2	−3	9.1	18	3.2	
25	3chy	_:57	_:65	1.5	2.9	31	2.1	0.4	1.1	−9	9.9	0.9	1.7	−13	24.0	9	3	
26	2hbg	_:18	_:26	5.9	6.9	51	0.8	3.3	4.4	45	4.5	0.9	1.2	26	7.9	10	10.7	
27	1gky	_:6	_:14	1.6	2.3	−6	1.6	1.0	1.3	−9	10.7	0.9	1.3	−15	10.7	20	3	
28	1ede	_:257	_:265	3.5	5.0	25	4.7	2.6	3.7	15	19.9	1.1	1.7	10	19.9	10		
29	3tgl	_:56	_:64	0.6	1.2	1	5.6	1.5	2.2	−4	10.7	1.5	2.2	−4	28.5	14		
30	2cyp	_:145	_:153	2.6	3.8	6	1.6	3.0	4.3	6	8.5	1.6	3.3	−1	11.7	29	3.5	
31	1gpr	_:63	_:71	1.7	4.0	−6	3.0	1.7	4.0	−6	17.6	1.7	4.0	−6	17.6	9		0.71
32	1ivd	_:244	_:252	1.3	2.1	−17	1.3	1.6	3.0	−20	5.8	1.7	3.0	−23	5.8	13	5.4	0.72
33	1gca	_:9	_:17	2.5	4.9	19	4.9	2.5	4.9	19	30.1	2.2	5.3	13	61.0	11	2.1	
34	1ptf	_:10	_:18	3.4	5.4	32	1.9	2.5	3.9	10	9.4	2.3	3.7	7	9.4	12		
35	2ayh	_:41	_:49	2.4	3.1	2	0.9	3.6	4.3	−8	6.0	3.2	4.1	−10	15.8	12	2.3	
36	1xif	_:59	_:67	2.9	3.2	9	1.4	3.2	3.5	−1	7.5	3.5	3.7	−7	20.7	26	14.1	
37	1tib	_:69	_:77	3.5	4.3	66	4.4	3.5	4.3	66	13.3	3.5	4.3	66	24.4	17		0.65
38	2cmd	_:81	_:89	3.8	5.1	15	2.1	4.3	5.7	0	11.7	4.5	5.6	−3	16.3	16	1.7	
39	1fus	_:31	_:39	2.5	3.9	100	1.9	5.3	5.7	100	12.9	4.5	5.2	46	17.2	13		
40	1byb	_:246	_:254	4.1	6.7	17	2.5	5.8	7.5	10	15.3	5.7	7.3	10	27.2	20	16.8	0.71
		AVERAGE		2.0	3.03	13	2.5	1.47	2.3	2	13.2	1.2	2.1	0	22.9			
		MEDIAN		1.6	3.01	11	2.0	0.65	1.5	1	9.9	0.6	1.2	−3	17.6			
		SALI GOOD			40%				68%				75%					
		SALI MEDIUM			50%				20%				15%					
		SALI BAD			10%				13%				10%					

All results are included (i.e., these are "unfiltered" results). For each stage in the hierarchical prediction procedure, the backbone and all-heavy atom (including side-chain) RMSDs are given for the lowest energy loops. The total computational time up to that stage, in CPU hours, is also provided, as well as the energy gap (dE, in kcal/mol), defined as the difference in energy between the lowest energy loop generated and the side-chain optimized/minimized native loop. The average B-factor for the backbone of the loop is also provided, as well as the smallest distance to the nearest ligand, if applicable, and the smallest "overlap factor" involving loop atoms in the native structure. Average and median RMSDs are provided for the entire test set, as well as the "good/medium/bad" classification, calculated in exactly the same manner as in the Fiser et al.[9] article. Note the substantial improvement in the overall results due to the refinement stages (2 and 3). The results are ordered according to the backbone RMSD of the lowest energy loop generated.

TABLE IX. Median Global Backbone RMSDs as a Function of Stage Number for the Combined/Full Data Set

	4	5	6	7	8	9	10	11	12
Stage 1	0.21	0.26	0.33	0.31	0.74	1.45	1.56	2.96	3.61
Stage 2	0.20	0.25	0.28	0.30	0.47	0.56	1.25	2.02	2.56
Stage 3	0.20	0.24	0.28	0.30	0.44	0.51	1.09	1.87	1.93

the scoring function. Whether a term of this type would improve results when a more accurate energy function is used remains to be determined.

A number of observations concerning the results of Xiang et al.[8] can be made by examining the detailed results for eight-residue loops provided in that article. First, if the test cases removed by our filtering procedure are also removed from the Xiang et al. results, the average RMSD changes minimally (from 1.45 Å to 1.49 Å), suggesting that the errors generated by ligands, ions, and abnormal pH are not dominant ones in their simulation protocol. Second, there is little difference between the median (1.33 Å) and average RMSDs in their results, indicating that errors are not concentrated in a small number of outliers (and hence localized to particular regions of phase space in the model), but rather are more or less uniformly distributed. This is unsurprising given the uncontrolled approximations that comprise their energy function.

Effects of Crystal Packing

We have performed a preliminary study of the role of crystal packing in determining loop conformations in our test set (i.e., by performing predictions on single proteins rather than in the crystal environment). The average RMSD for short loops appears to be nearly identical with/without crystal packing. For the Fiser et al.[9] four-residue loop test set, the average backbone RMSDs are 0.23/0.28 Å with/without crystal packing, and for six-residue loops the values are 0.68/0.61 Å (the slightly better RMSD without crystal packing is probably just statistical noise). Maximal differences are observed for eight-residue loops, with the average backbone RMSD increasing substantially from 1.05 to 1.46 Å when crystal-packing forces are removed. For the Fiser et al.[9] 10-residue loops, the average RMSD is the same with and without crystal packing, 1.90 Å. This should not be taken to imply that crystal-packing forces are unimportant for longer loops. On the contrary, although it makes sense that short loops should be relatively unaffected, we see no reason why long loops should be. Rather, we expect that with complete sampling and a perfect energy function, significant crystal-packing effects would be observed; errors due to incomplete sampling and an imperfect energy function are simply similar or larger in magnitude than errors due to neglect of crystal packing for the longer loops.

Computational Effort

Figure 7 presents the average CPU time as a function of loop length, averaged over all of the test cases. Calculations were conducted by running on a single node of Intel or AMD PCs operating in the range of 1.4 GHz. There is some difference in the performance of the various PCs (on the order of 30%), but not enough to affect the qualitative analysis of the computational requirements of this methodology, which is the objective of this section.

There is a considerable variance in CPU time for a given loop length, dependent on a variety of factors (e.g., steric restrictions on the loop). The maximum time obtained for any loop was approximately 160 h for one of the 12-residue loops. The scaling of total CPU time with loop length is approximately cubic. In most cases, the computational time is dominated by side-chain optimization and minimization on the loop candidates chosen from clustering, although for the longer loops, particularly "floppy" ones, the initial loop buildup can also constitute a nonnegligible fraction of the computational time (other portions of the algorithm, such as clustering, contribute only negligibly). The number of loop conformations subjected to side-chain optimization and minimization is chosen to scale linearly with loop length. Thus, the computational expense associated with side-chain optimization and minimization can be surmised to scale approximately quadratically with loop length. Although linear scaling would of course be preferable, we see no way to obtain such scaling without sacrificing considerable accuracy. Polynomial scaling of course is vastly preferable to exponential scaling, which would be expected for any naive buildup algorithm.

For prediction of hundreds or even thousands of individual loop regions, these CPU times are unproblematic given the low cost of PC processors. However, realistic homology modeling may require repeated predictions of individual regions to obtain an improved overall structure, because of couplings between different regions of the protein (one also has to allow the side-chains in the body of the protein to relax; preliminary results indicate that reasonable prediction accuracy can be maintained in this situation, but at a CPU cost of 2–3 times what we report here for a rigid protein). Although a rigorous analysis of how effective the approach presented in this article will be for such homology modeling efforts (which in most cases will be restricted to the active site, an important factor in reducing computational effort) awaits explicit studies, our best guess is that these calculations will be tractable for an individual protein on a modest sized PC cluster (~16–32 nodes) using distributed computing methods. This is, of course, considerably more computational effort than is required for simple model-building programs or database type approximations to loop structure and is likely not suitable for application on a genomic scale (although this is not inconceivable with a sufficiently large cluster). On the other hand, our expectation is that the accuracy will be substantially improved compared with simpler approaches.

This trade-off of speed versus accuracy is typical as one increases the realism of the model; the present algorithm, because of its hierarchical nature, for the first time makes it feasible to consider the use of a model based on accurate physical principles in practically addressing high-resolution protein structure refinement.

CONCLUSION

The results discussed above show a qualitative improvement in loop prediction accuracy compared with previous methods in the literature. We attribute this success primarily to the use of an all-atom protein force field and, especially, realistic treatment of solvent, which is clearly a necessity for properly describing structures and energies on the protein surface. Although the use of such an all-atom, physics-based energy function can be expected to improve accuracy relative to prior approaches, most of which have used simplified or heuristic scoring functions, the associated energy surface is extremely rugged and poses severe challenges for conformational sampling. In particular, conformational sampling of the loop backbone must be coupled with extensive sampling of the loop side-chains. We have addressed this challenge through the development of a series of new algorithms that exploit hierarchical and multiscale concepts. In the tests reported here, the sampling algorithm identifies local minima with energies that are lower than or essentially equal to the minimized native loop the vast majority of the time. In this sense, the accuracy of the predictions is limited primarily by the accuracy of the energy function, and further improvements to the sampling are unlikely to strongly improve accuracy. Sampling remains challenging for loops longer than 12 residues (and especially for those longer than 15 residues), and continued algorithmic development is underway for these cases. At all loop lengths, further improvements in computational efficiency are clearly achievable and are being actively pursued.

The protein molecular mechanics energy function and continuum solvation model used here perform at the state of the art, and this is sufficient to obtain remarkably good structural predictions—in many cases, better than 0.5 Å—in a completely ab initio fashion. In cases where incorrect loop conformations are generated with much lower energies than the minimized native, the energy function itself is not necessarily to blame. For example, we have identified a number of cases where relatively low-quality structures (very high B-factors or steric clashes in the loop) are likely to be responsible for the native structure not receiving the lowest energy. We have also made only a rudimentary effort to assign protonation states to the titratable side-chains of the protein (based solely on pH and not on the environment of the side-chains), and protonation states are considered fixed. We believe, but cannot prove at this time, that errors in protonation states are responsible for a large fraction of the remaining prediction errors and are pursuing strategies to efficiently couple protonation state prediction with loop conformational sampling. Internal entropy contributions to the free energy have also been neglected here, and this omission could, in principle, affect accuracy for the small minority of cases that exhibit evidence of multiple nearly degenerate energy basins. Despite these caveats, detailed analysis of the incorrect predictions is underway to identify cases where the energy function itself is likely to be limiting the accuracy and, thus, to motivate possible strategies for improving the energy function.

One potential application of our algorithm is to the refinement of protein models from X-ray crystallography, as well as structure solution by molecular replacement. Ultimately, the major application of our loop prediction algorithm is to the refinement of protein homology models, and several challenges remain to be addressed for this application. First, in many homology-modeling applications, the conformations of the side-chains surrounding the loop to be predicted are not known with high accuracy and, thus, must be predicted concurrently with the loop itself. Initial results indicate that it is possible to retain excellent loop prediction accuracy when side-chains in the protein native structures are perturbed from their original positions, an essential property to ensure usefulness for the homology-modeling problem. Our strategy simply involves sampling side-chains located spatially close to the loop at the same time that the side-chains on the loop itself are sampled (i.e., after loop candidate generation, screening, and clustering). We have evaluated the effectiveness of this strategy in a controlled way by randomly perturbing loops from the native conformation (typical RMSD is 3.0 Å) and then performing side-chain optimization on the full protein in the new conformation. In initial tests on the Fiser et al.[9] seven-residue loop test set, the average backbone RMSD increases from 0.51 Å for native loop reconstruction to 0.71 Å for the loop prediction stating from non-native structures. It should be noted that crystal-packing effects are not taken into account in the latter calculation and that this omission accounts for a large portion of the increase in average RMSD; native loop reconstruction without crystal packing yields an average RMSD of 0.65 Å.

Additional problems related to loop optimization in realistic homology-modeling scenarios include the simultaneous refinement of the positions of multiple loops in close proximity and dealing with imperfect backbone structures (i.e., outside the loop region; this topic has been explored previously by Fiser et al.[9]). We have begun to develop an iterative approach in which local optimizations (i.e., of individual loops, and also rigid body motions of helices[36]) are applied iteratively to a protein model until the all-atom energy stops decreasing significantly. An early version of this strategy was used in the recent CASP5 contest. The loop prediction algorithm contributed to successful refinements (i.e., reduction of overall backbone errors, as measured by the GDT-TS score used in the CASP assessment) of several models in the "homology" category (i.e., those targets for which sequence alignments are less difficult, especially T133, T150, T178, and T186; these results are reported elsewhere).[37]

ADDITIONAL INFORMATION

Numerous additional tables are presented in Supplementary Materials, including a detailed analysis for every loop in the test set. The loop prediction algorithm is implemented as part of the Protein Local Optimization Program. Information about obtaining the program is available at the corresponding author's Web site (http://francisco.compchem.ucsf.edu/~jacobson/), as well as a user manual. On this Web site, we have also made available loop "decoy sets" generated from the results described here. Specifically, we have concatenated all loop conformations that were subjected to side-chain optimization and minimization at any stage of the hierarchical prediction algorithm.

ACKNOWLEDGMENTS

MPJ acknowledges prior support from a National Science Foundation Postdoctoral Fellowship in Bioinformatics, and current support from National Science Foundation. We thank our collaborator Dr. Zhexin Xiang (Columbia University Department of Biochemistry and Biophysics) for sharing his side-chain rotamer libraries with us, for many helpful discussions, and for sharing his own loop prediction results in advance of publication. We also thank Prof. Ron Levy (Rutgers) and his group for sharing results on database approaches to loop prediction and for testing early versions of our code.

REFERENCES

1. 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.
2. Rufino SD, Donate LE, Canard LHJ, Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: application to comparative modelling. J Mol Biol 1997;267:352–367.
3. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJE. An automated classification of the structure of protein loops. J Mol Biol 1997;266:814–830.
4. Wojcik J, Mornon JP, Chomilier J. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. J Mol Biol 1999;289:1469–1490.
5. vanVlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
6. Moult J, James MNG. An algorithm which predicts the conformation of short lengths of chain in proteins. J Mol Graph 1986;4:180–180.
7. Shenkin PS, Yarmush DL, Fine RM, Levinthal C. Method for quickly generating random conformations of ring-like structures for subsequent energy minimization or molecular-dynamics—application to antibody hypervariable loops. Biophys J 1987;51:A232–A232.
8. 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.
9. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
10. Galaktionov S, Nikiforovich GV, Marshall GR. Ab initio modeling of small, medium, and large loops in proteins. Biopolymers 2001;60:153–168.
11. Dudek MJ, Scheraga HA. Protein-structure prediction using a combination of sequence homology and global energy minimization .1. Global energy minimization of surface loops. J Comput Chem 1990;11:121–151.
12. Palmer KA, Scheraga HA. Standard-geometry chains fitted to x-ray derived structures—validation of the rigid-geometry approximation. 2. Systematic searches for short loops in proteins—applications to bovine pancreatic ribonuclease-a and human lysozyme. J Comput Chem 1992;13:329–350.
13. Das B, Meirovitch H. Optimization of solvation models for predicting the structure of surface loops in proteins. Proteins 2001;43:303–314.
14. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
15. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized born model of solvation effects. Proteins 1999;35:173–183.
16. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
17. DePristo MA, de Bakker PIW, Lovell SC, Blundell TL. Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 2003;51:41–55.
18. Zhang H, Lai L., Wang L, Han Y, Tang Y. A fast and efficient program for modeling protein loops. Biopolymers 1997;41:61–72.
19. Wedemeyer WJ, Scheraga HA. Exact analytical loop closure in proteins using polynomial equations. J Comput Chem 1999;20:819–844.
20. Go N, Scheraga HA. Ring closure and local conformational deformations of chain molecules. Biopolymers 1970;3:178–187.
21. Bruccoleri RE, Karplus M. Chain closure with bond angle variations. Macromolecules 1985;18:2767–2773.
22. Smith KC, Honig B. Evaluation of the conformational free-energies of loops in proteins. Proteins 1994;18:119–132.
23. 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 2003;51:21–40.
24. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002;106:11673–11680.
25. 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.
26. 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;105:6474–6487.
27. 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 estimators. J Comput Chem 2002;23:517–529.
28. Ghosh A, Rapp CS, Friesner RA. Generalized born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.
29. Hartigan JA. Clustering algorithms. New York: Wiley; 1975.
30. Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. Appl Stat 1979;28:100–108.
31. Xiang ZX, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
32. Xie DX, Schlick T. Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications. Siam J Optimiz 1999;10:132–154.
33. Schlick T, Overton M. A powerful truncated Newton method for potential-energy minimization. J Comput Chem 1987;8:1025–1039.
34. Schlick T, Fogelson A. Algorithm 702—Tnpack—a truncated Newton minimization package for large-scale problems. 1. Algorithm and usage. Acm T Math Software 1992;18:141–141.
35. Rakhmanov EA, Saff EB, Zhou YM. Minimal discrete energy on the sphere. Math Res Lett 1994;1:674–662.
36. Li X, Jacobson MP, Friesner RA. High resolution prediction of protein helix positions and orientations. Proteins 2003. Accepted for publication.
37. Jacobson MP, Pincus DL, Day TJF, Rapp CS, Li X, An Y, Friesner RA. Use of all-atom physical chemistry energy functions for comparative model construction, selection, and refinement. Protein Sci 2003. Submitted for publication.