High-Resolution Prediction of Protein Helix Positions and Orientations

Xin Li,[1] Matthew P. Jacobson,[2*] and Richard A. Friesner[1]
[1]*Department of Chemistry, Columbia University, New York, New York*
[2]*Department of Pharmaceutical Chemistry, University of California, San Francisco, California*

ABSTRACT **We have developed a new method for predicting helix positions in globular proteins that is intended primarily for comparative modeling and other applications where high precision is required. Unlike helix packing algorithms designed for ab initio folding, we assume that knowledge is available about the qualitative placement of all helices. However, even among homologous proteins, the corresponding helices can demonstrate substantial differences in positions and orientations, and for this reason, improperly positioned helices can contribute significantly to the overall backbone root-mean-square deviation (RMSD) of comparative models. A helix packing algorithm for use in comparative modeling must obtain high precision to be useful, and for this reason we utilize an all-atom protein force field (OPLS) and a Generalized Born continuum solvent model. To reduce the computational expense associated with using a detailed, physics-based energy function, we have developed new hierarchical and multiscale algorithms for sampling the helices and flanking loops. We validate the method using a test suite of 33 cases, which are drawn from a diverse set of high-resolution crystal structures. The helix positions are reproduced with an average backbone RMSD of 0.6 Å, while the average backbone RMSD of the complete loop–helix–loop region (i.e., the helix with the surrounding loops, which are also repredicted) is 1.3 Å. Proteins 2004;55:368–382.** © **2004 Wiley-Liss, Inc.**

Key words: homology modeling; helix prediction; all-atom force field; conformational sampling; generalized Born

INTRODUCTION

Several lines of research have investigated the packing and dynamics of helices in proteins, including the following:

1. *Molecular dynamics (MD) investigations of fluctuations in helix positions.*[1–4] Elber and Karplus studied the helix movement of myoglobin through an MD simulation.[2] Later, Rojewska and Elber performed another MD study on the secondary structures of proteins, in which they chose myohemerythrin as their case of study.[3] Furois-Corbin et al. studied motions of helices in deoxymyoglobin.[4] This line of research has emphasized that many helices move as semirigid objects, in the sense that backbone fluctuations internal to the helix are small, but the helices can move relative to the remainder of the protein by 1–2 Å on a picosecond timescale.
2. *Modeling of membrane proteins.*[5–9] The small number of available structures of membrane proteins and the importance of transmembrane helices in the structure and function of many membrane proteins have motivated the development of algorithms suitable for prediction of helix packing arrangements in the membrane environment.
3. *Helix packing algorithms for "ab initio" protein structure prediction.*[10–15] Many groups, including our own, have developed algorithms for predicting the structure of helical domains, typically via Monte Carlo sampling using a reduced model potential function. Two recent contributions are from Fain and Levitt[10] and Nanias et al.,[11] who described a global optimization procedure for helix packing that utilized a simplified Lennard–Jones potential energy function.
4. *Helix movements associated with ligand binding.*[16–18] Kamiya and Reynolds have postulated, on the basis of computational modeling, that lateral helix motions are required for binding of certain ligands to β-andrenergic receptor, a G-protein coupled receptor.[16] Several lines of evidence suggest that a ligand-induced, piston-like sliding of a long helix is responsible for signal transmission in bacterial chemoreceptors.[17] The binding of agonists and antagonists to estrogen receptor trigger distinct conformational changes, dominated by movement of helix 12.[18]
5. *Comparative analysis of helix conformations in related protein structures.*[19–22] Chothia et al.[19,21] inferred from multiple, independent crystal structures of insulin that

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: Prior support to M. P. Jacobson from an National Science Foundation Postdoctoral Fellowship in Bioinformatics, and current support from NSF; Grant number: 0302445. Grant sponsor: National Institutes of Health; Grant number: GM-52018 to R.A. Friesner.

*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 29 June 2003; Accepted 28 September 2003

Published online 5 March 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20014

helices can undergo low-energy shifts of up to ~1.5 Å. Chothia and Lesk[20] also studied the role of helix movements in reconstructing the heme pocket during the evolution of the cytochrome c family, and Gerstein et al.[22] implicated concerted helix shifts as a mechanism for protein domain movement.

One common feature of these diverse lines of research is that helices are implicitly or explicitly considered to move as rigid objects, an assumption that also underlies the work described here.

Our own work is motivated primarily by the homology modeling problem. In most homology modeling protocols, α-helices, like β-sheets, are typically treated as "conserved" secondary structure elements; that is, during construction and refinement of homology models, computational effort is directed primarily toward side-chain and loop prediction, with little deliberate sampling of the remainder of the protein model. However, in homologous proteins, the corresponding aligned helices can demonstrate substantial differences in positions and orientations, as measured via backbone root-mean-square deviation (RMSD). One example is the target T0122 in CASP4.[23] The sequence identity between this target [Protein Data Bank (PDB) code: 1geq] and its closest template (PDB code: 1qop) is 33%. However, a structural alignment between these two proteins yields an RMSD of 4.1 Å for a 10-residue α-helix ranging from residues 98 to 108, which represents a major contribution to the overall backbone RMSD of 1.7 Å. In our own participation in CASP5, we demonstrated that it is possible to improve the quality of comparative models using a strategy that includes an early version of the helix prediction algorithm described here. In particular, helix repacking contributed to lowering the backbone RMSD of our models for targets T132, T133, T150, and T178, a result that is reported elsewhere.[24]

In contrast to a significant amount of literature on loop prediction for purposes of homology modeling,[25–41] most algorithms for helix packing have been directed toward ab initio folding[10–14] or modeling of membrane proteins,[5–9] as discussed above. In this article, we introduce a methodology for predicting the conformations of helices and their flanking loops that is designed for comparative modeling applications (i.e., when knowledge can be assumed about the qualitative placement of all helices, but details of their packing are unknown and require prediction). A helix packing algorithm for use in comparative modeling must obtain high precision to be useful, and for this reason we utilize an all-atom protein force field (OPLS[42–44]) and a continuum solvation model (Surface Generalized Born, SGB[42]). We have invested substantial effort in developing hierarchical and multiscale algorithms to reduce the computational expense associated with using a detailed, physics-based energy function. A significant number of techniques are borrowed from a previously described loop prediction method,[25] such as iterative sampling and elimination of redundant structures via clustering,[45–47] fast algorithms for side-chain optimization,[48–50] and complete energy minimization on the probed region using a novel, multiscale minimization algorithm (Jacobson and Friesner, unpublished results). In addition, the methods for predicting conformations of the loops flanking the helices, which is an integral part of the prediction methodology, are borrowed intact from the loop refinement methods,[25] other than some minor modifications.

However, we have also developed new algorithms expressly designed for manipulation of the helix itself. A key hypothesis underlying our algorithm is that helix packing with the body of the protein dominates the interaction energy, as opposed to the flanking loops. This suggests an approach in which the helix is initially docked (eliminating positions where the addition of the flanking loops would be impossible based on simple geometrical or steric criteria), the best scoring helix positions are selected, and the flanking loops are then optimized for these helix positions. To implement this strategy, we have devised efficient techniques for sampling the helix positions and orientations in a systematic, exhaustive way, through a three-dimensional (3D) grid search. The resulting helix conformational space is then pruned by various geometrically or physically based screens to remove unreasonable helix conformations.

The algorithm is described in detail in the Methods section, and its validation is reported in the Results and Discussion section. Specifically, the algorithm is validated in a manner analogous to most tests of loop prediction algorithms: The conformation of a helix and its surrounding loops is predicted, holding the remainder of the protein fixed at its experimentally determined structure. We utilize a test suite of 33 cases, which are drawn from a diverse set of high-resolution crystal structures. The results are generally quite satisfactory. The helix positions are reproduced with an average backbone RMSD of 0.6 Å, while the average backbone RMSD of the complete loop–helix–loop region (i.e., the helix with the surrounding loops, which are also repredicted) is 1.25 Å.

METHODS

Overview

The loop–helix–loop prediction problem can be conceptualized as the prediction of a long loop with a rigid segment (the helix) in the middle. The helix itself is allowed 6 degrees of freedom: 3 translations and 3 rotations. Thus, one might naively expect the computational complexity to be somewhat greater than that associated with predicting a single loop with the number of residues equal to the number of residues in the loops flanking the helix. In this work, we sample up to 8 residues in each loop flanking the helix in question; this would imply that the computational expense should be greater than that associated with sampling a 16-residue loop, which is extremely challenging. In practice, we reduce the computational expense in several ways, including the extensive use of hierarchical algorithms and clustering, as described below. Most importantly, we postulate that the energy of the loop–helix–loop region interacting with the remainder of the protein will be dominated by the helix region. This suggests that we can identify favorable locations for the helix prior to sampling the loops at all. In addition, the 2 loops rarely interact strongly with each other, implying that they can be sampled independently. The high accuracy achieved by our algorithm suggests that these hypotheses are reasonable.

Our helix prediction algorithm proceeds through the following 5 steps:

1. Initial sampling of helix positions, using geometric transformations to generate many candidates, followed by multiple stages of screening to eliminate conformations based on a variety of geometrically and physically based criteria.
2. Clustering[46,47] of the surviving helix conformations to reduce redundancy, and selection of representative candidates from each cluster.
3. A coarse-grained loop construction procedure, to quickly close the two flanking loops on either side of the helix.
4. A multistage, hierarchical approach to refining the conformations of the flanking loops, as described in an earlier publication of our group.[25]
5. Side-chain optimization[48,50] and complete energy minimization of the entire loop–helix–loop region.

In implementation, our strategy basically contains two separate stages: helix sampling (steps 1–3) and flanking loop refinement (steps 4 and 5). The helix sampling stage generates a set of low-energy helix positions by geometric enumeration and screenings, followed by clustering and a fast loop closure procedure. This stage focuses on the sampling and scoring of the helix body itself, and pays less attention to the flanking loops beyond ensuring that they are in geometrically reasonable conformations (no steric clashes or Ramachandran violations) at the end of the helix sampling stage. Energy-based refinement is then applied to the two flanking loops using a hierarchical, multistage algorithm.[25] This stage often results in a significant decrease in the total all-atom energy, as well as the RMSDs for the two loop regions (and, in many cases, of the helix itself, either via motion of the helix in the final energy minimization or by selection of a superior helix position due to improvement of the energy of the various candidates passed into the refinement stage).

Energy Function

All energy evaluations are performed using an all-atom energy model, based upon the OPLS-AA force field[43,44] and the SGB implicit solvent model.[42] The force field and solvation model have been extensively discussed and tested in previous publications (including their effectiveness in loop and side-chain prediction),[25,49] and we do not repeat this discussion here. The results obtained below provide further assessment of the quality of the potential energy functions and solvation model. As in previous work, cases in which structures with high RMSD from the native are found at significantly lower energies than the native may indicate a problem with the energy model, and will be used in future efforts to develop and validate improvements in the force field and solvation terms.

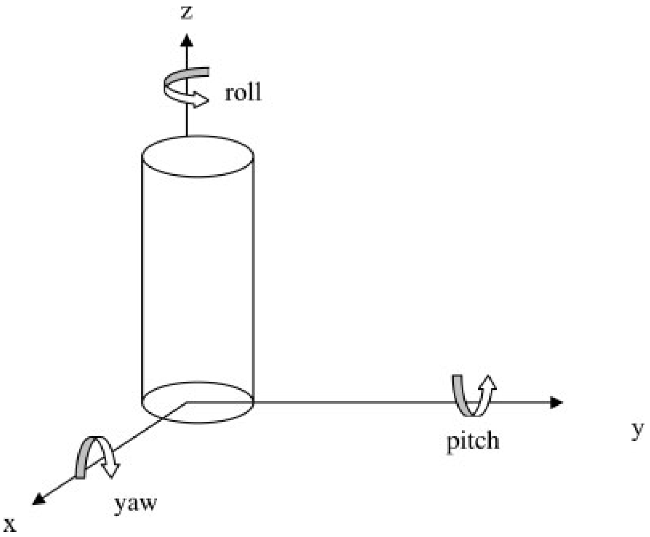

Fig. 1. Geometric model of the α-helix, illustrating the angular degrees of freedom used in the sampling algorithm.

Step 1: Enumeration and Screenings Of Helix Conformations

In this work, we restrict ourselves to sampling of a single helix in a globular protein. The definition and calculation of the helix axis are based on an algorithm proposed earlier by Christopher et al.[51] The helix is fitted into a geometrically "ideal" and rigid helix body, and its axis is calculated.[51] In this work, the backbone conformation of the helix is not modified from the conformation found in the native protein. We then sample positions of the helix by moving it as a rigid body using all 6 degrees of freedom (Fig. 1). Three of the 6 degrees of freedom are the "translational" degrees of freedom [i.e., the helix center can move in the longitudinal direction along the helix axis (z axis), or in two mutually orthogonal directions perpendicular to the helix axis (x and y axes)]. Two of the angular degrees of freedom are related to "tilting" of the helix; that is, the helix can rotate along the two axes orthogonal to the helix axis (x and y axes) with the helix center kept fixed. The final degree of freedom specifies rotation about the helix axis itself (which we will also refer to as the "roll" degree of freedom). Mathematically, our description of the angular degrees of freedom is similar to the so-called Euler angle description, although in practice we use the equivalent yaw–pitch–roll convention.

We initially enumerate helix positions by a novel grid search algorithm (Fig. 2). We first temporarily remove the two flanking loops, leaving the helix "floating in the air," and then set up two bounding spheres centered on the two end points of the loop–helix–loop region. The radii of the bounding spheres are chosen to ensure that the helix cannot move so far away that the flanking loops will fail to close in later stages. We then define a set of grid points inside both spheres based on the sampling grid resolution, which is a user-adjustable parameter. Finer resolution will, of course, result in more accuracy but will also lead to more candidate helix positions (which scale approximately

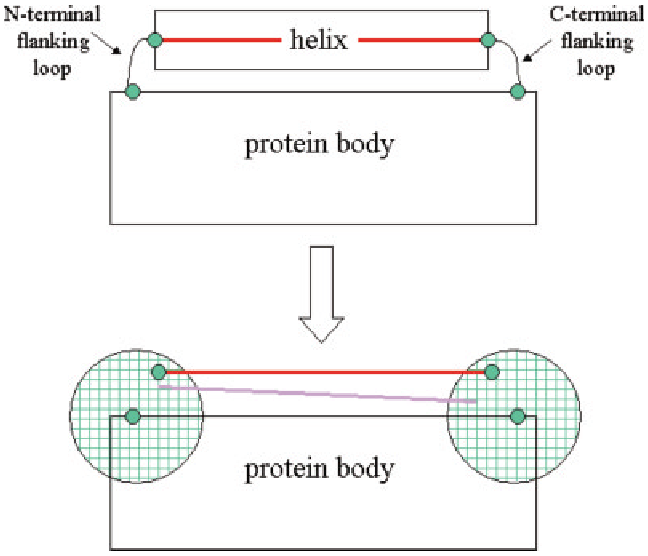

Fig. 2. Two-dimensional illustration of the grid-based procedure for exhaustively generating helix positions. Spheres are centered on the points of attachment between the flanking loops and the body of the protein, with radii determined by the lengths of the flanking loops; that is, the spheres represent all possible points of attachment between the flanking loops and the helix. The volumes of the spheres are filled with a 3D grid, with a resolution, typically, of 1 Å (the precise resolution is adaptively adjusted in practice; see text for details). Note that many grid points can be immediately rejected on the basis that they lie inside the protein "body." All pairs of grid opints, one in each sphere, that satisfy the helix length constraint are retained. The red bar represents the correct helix axis, and the purple bar is an example of an incorrect one.

as the sixth power of the number of grid points). For our purpose, we take the starting grid resolution to be 1 Å. Our program adaptively adjusts the sampling resolution should it generate either too many (>500) or too few (<60) conformations, after all of the screens have been applied, as discussed below. The resolution is adjusted in increments of 0.25 Å until the number of accepted helix conformations falls between the two limits.

We then identify all pairs of points—one point in each sphere—that are separated by a distance nearly equal to the length of the helix. In practice, deviations of up to 1% are tolerated; approximately 1% of the possible pairs of points satisfy this criterion. A straight line joining each retained pair of grid points uniquely defines a possible position and orientation of the helix axis. However, specification of the helix axis only determines 5 of the 6 degrees of freedom, which is not sufficient to uniquely determine the conformation of the whole helix. The final degree of freedom, the "roll" angle degree of freedom, is still arbitrary, as rotation along the helix axis will leave both the helix center and the helix axis invariant. We therefore introduce another user-adjustable parameter called d_{roll}, the sampling resolution for the roll degree of freedom. In the results reported here, we use $d_{roll} = 15°$. To avoid any bias arising from prior knowledge of the native roll angle, we deliberately add an artificial "offset" equal to $d_{roll}/2$ as our initial position of rotation; that is, we start our sampling with the helix rotated 7.5° from the native position, rather than 0°. This represents the worst-case

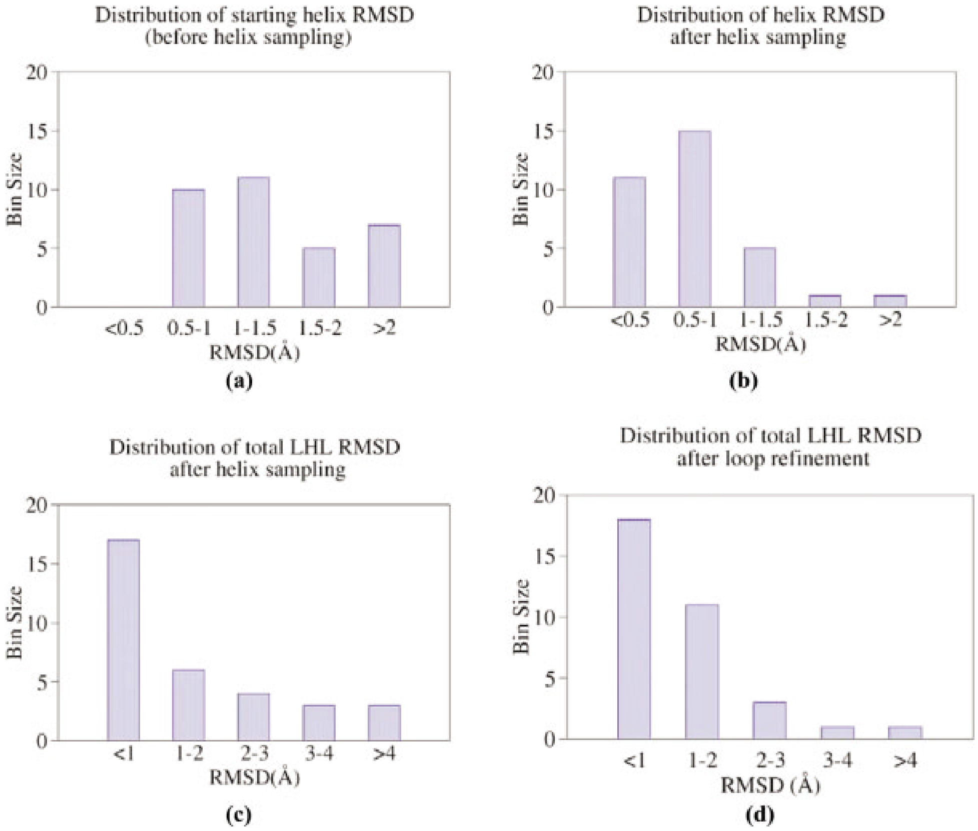

Fig. 3. Distribution of RMSDs among all test cases at various stages of the prediction algorithm. (**a**) Distribution of starting helix RMSD (before helix sampling). (**b**) Distribution of helix RMSD after helix sampling. (**c**) Distribution of total loop–helix–loop (LHL) RMSD after helix sampling. (**d**) Distribution of total LHL RMSD after loop refinement.

scenario as compared to a random initial roll position; hence, performance in a blind test should be at least as good as what we report here.

With the number of grid points inside the spheres at the two ends of the helix being N_1 and N_2, respectively, the total number of potential helix conformations is N_{rot} = $(360/d_{roll})$ $*N_1*N_2$. This is typically a very large number, on the order of 10^6, with 1 Å grid resolution. The vast majority of these conformations are eliminated on the basis of the helix length restraint discussed above. Many of the remaining conformations can be eliminated by other rapid screens that filter out clearly unreasonable helix positions, resulting in a significant reduction in the number of candidate helix conformations passed onto the loop refinement stage. The screening stage in fact is critical in making our helix sampling algorithm a tractable computational method. These screenings include the following:

1. *Loop length.* This screen ensures that the helix cannot assume a position that will make it impossible to close the flanking loops. We use a simple distance criterion (i.e., between the end of the helix and the point at which the flanking loop connects to the body of the protein) based on statistical analysis of end-to-end distances of loop regions in a test set of over 500 proteins.[25] At the 99.5% level, the maximum distance that can be spanned by 4 residues (C^{α}–C^{α} distance) is 13.97 Å, for example.
2. *Side-chain clashes.* This screen is designed to ensure that sufficient space exists for the side-chains on the helix. We eliminate those conformations that will result in unavoidable clashes between one or more side-chains on the helix and the remainder of the protein (excluding the flanking loops, which have not been rebuilt). We utilize a 30° side-chain rotamer library for this task, and a helix conformation is judged acceptable if at least one conformation can be found for each side-chain that is free from steric clashes. This screen does not guarantee that it is possible to pack the side-chains of the helix together in a combinatorial sense, but does eliminate large numbers of incorrect helix conformations.
3. *Distance of helix from protein body.* This screen ensures that the helix will not lie either too close or too far away from the remainder of the protein (again, excluding the flanking loops). This is enforced using the closest distances between any atom of the rest of the protein body to 3 points on the helix axis: the 2 end points and the midpoint. The minimum/maximum allowable distances for the 2 helix end points are 4.0 and 11.0 Å, and for the midpoint are 5.5 and 10.0 Å. These values were obtained empirically by examining a diverse set of helices in high-resolution crystal structures.

In addition, we introduce another parameter, *max_dev*, that prevents the helix from moving too far way from its starting position, which in this case is the native geometry. Helix moves resulting in conformations with either end point deviating larger than *max_dev* from its original position will be discarded. The value of *max_dev* is 5 Å in the results discussed below, although it is user-adjustable. The rationale for this restriction on the helix sampling is that one would not typically attempt to move a helix by more than this amount in most comparative modeling applications. Initial evaluation of homologous pairs of proteins with nontrivial differences in the conformations of corresponding helices indicates that a 5-Å restriction on each helix end point is sufficient to allow the correct solution to be generated in the great majority of homology modeling applications. Larger shifts can be allowed (as might be appropriate in a fold recognition context, or in a case where the helix appears to be an insertion from the point of view of the template), albeit at the cost of increased computational expense. Application to such cases will be explored in future publications; here, we focus on the more modest sampling required by more typical comparative modeling applications.

Step 2: Clustering of Helix Conformations

We utilize the K-means clustering algorithm[46,47] to reduce the redundancy in the set of helix conformations resulting from step 1. The descriptors used for clustering are the Cartesian coordinates of the two end points of the helix axis, plus the sine and cosine of the "roll" rotation angle (both are needed to account for the 360° periodicity), resulting in a total of 8 descriptors. In the work described here, the number of clusters is set as twice the number of residues of the entire loop–helix–loop region. Although the number of clusters must be chosen in advance, we have also implemented a simple method of adaptively increasing the number of clusters when 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 2 new clusters, and the clustering is repeated.

To choose helix conformations for further optimization, a "representative" member of each cluster is chosen, as the member of the cluster located closest to the centroid of the cluster (i.e., in the 8D space of the descriptors). If this representative should fail at a later stage of optimization (e.g., it is found to be impossible to close the flanking loops), then optimization of other members of the cluster are attempted, again using the most "central" remaining member first.

Step 3: Loop Closure and Screening

At this point, only the helix conformation is specified, and the flanking loops must be predicted. We initially attempt to find any closed loop conformation that avoids steric clashes and utilizes only "Ramachandran-allowed" dihedral angles for the backbone; no energetic criteria are used at this stage (energy-based optimization is performed in step 4). The algorithm for generating loop conformations has been described earlier.[25] In brief, candidate backbone conformations are generated by splitting the loop in half and using a build-up procedure for each side, eventually joining the loops in the middle. The build-up algorithm employs a rotamer-like library of acceptable backbone dihedral angle combinations (ϕ,ψ) [i.e., based on the Ramachandran map (Pro and Gly are treated separately)]. During the build-up procedure, conformations are rejected (1) if steric clashes exist with the body of the protein, (2) if there is insufficient space for loop side-chains to fit adequately, (3) if the loop travels too far from the body of the protein, or (4) if the loop cannot close, based on geometric criteria. When all acceptable conformations of the C- and N-terminal halves of the loop have been identified, at a given sampling resolution, then closed loops are generated by pairing halves whose end points lie close in space. Resultant closed loops are rejected (1) if the backbone dihedrals of the closure residue lie outside Ramachandran-allowed regions, (2) if steric clashes exist between the two halves, and (3) if there is insufficient space for loop side-chains to fit without steric clashes. The sampling resolution for the backbone dihedral angles is gradually decreased (i.e., finer sampling) until a satisfactory closed loop conformation is generated. The search is terminated (and it is assumed that no satisfactory conformation is possible given the helix conformation) if at least 2500 conformations have been generated for each side of the optimization region at the build-up phase, but no satisfactory closed loops have been found.

Once both flanking loops are closed, the entire loop–helix–loop region is subjected to side-chain optimization and full energy minimization (see step 5). At the end of the loop closure step, the helix sampling stage is finished, and the loop–helix–loop conformations (one from each cluster obtained in step 2) are ranked in order of their all-atom energies.

Step 4: Multistage Iterative Loop Refinement on the Two Flanking Loops

At this point, the flanking loops are closed but not necessarily in energetically optimal conformations. For each target, the 5 loop–helix–loop conformations with the lowest all-atom energy at the end of the helix sampling stage are subjected to energy-based optimization of the loops. This is accomplished using the hierarchical loop refinement algorithm described previously.[25] The procedure for generating loop conformations is outlined in step 3. Instead of choosing any geometrically reasonable loop conformation, however, hundreds or thousands of loop candidates are generated and then subjected to clustering (similar to step 2 above), side-chain optimization, and complete energy minimization. This energy-based loop optimization algorithm is applied hierarchically and iteratively. The initial optimization stage consists of unconstrained loop generation to coarsely sample the conformational space. Two 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 loops on either end of the helix are refined independently and simultaneously, and the lowest energy loops from each simulation are combined into a single model.

Step 5: Side-Chain Optimization and Full Energy Minimization

After the loop refinement, the entire loop–helix–loop region is subjected to side-chain optimization and full energy minimization.

The side-chain optimization algorithm has been described previously.[50] Sampling is accomplished primarily by using a highly detailed (10° resolution) rotamer library constructed by Xiang and Honig[48]; this library contains, for example, $>$2000 rotamers for Arg and Lys. Prescreening the rotamers for steric clashes using geometric criteria prior to performing any energy evaluations reduces the computational expense. The combinatorial optimization algorithm is also adapted from the method of Xiang and Honig,[48] which is similar in spirit to earlier work by Bruccoleri and Karplus.[28] 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 self-consistency (i.e., side-chains cease changing rotamer states). Side-chain entropy is partially accounted for by enumerating side-chain conformations on a dense dihedral angle grid, and calculating the configuration integral for low-energy basins; this procedure does not, however, account for entropy associated with correlated motions of side-chains.

Our minimization algorithm is based upon the truncated Newton (TN) method, specifically the TNPACK implementation of Schlick and coworkers.[52–54] We have modified the algorithm to improve its efficiency in 2 ways. First, we have adopted multiscale concepts similar to those underlying MD methods such as RESPA[55]; that is, we divide the molecular mechanics forces into short (bond, angle, torsion, nonbonded interactions between atoms separated by $<$10 Å) and long-range components (all other forces), and update the short-range forces more frequently than the long-range forces. The minimization is considered to be converged when the residual root-mean-square gradient is $<$0.001 kcal/mol/Å.

Our second modification is a self-consistent procedure for minimization in Generalized Born (GB) solvent. In brief, the Born α's are held fixed during the course of a minimization, then updated prior to another minimization, and so on, until the energy ceases to decrease by more than 1 kcal/mol; that is, we exploit the slowly varying nature of the Born α's, as well as the analytical and differentiable pair screening term.[56] In practice, self-consistency rarely requires more than 2 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 $\sim$50% greater computational expense than vacuum minimizations.

For each target, the final structure with the lowest all-atom energy is then chosen as the best prediction.

Crystal Packing

In previous work,[50] we have found that crystal packing forces can affect structural details of proteins, especially the conformations of polar side-chains on the surfaces of proteins. To remove any uncertainty about effects of neglecting crystal packing, and to provide a fair comparison with experimental crystal structures, we perform all predictions in the crystal environment; that is, crystal unit cells are explicitly reconstructed using the dimensions and space group reported in the PDB files. We do not attempt to employ explicit lattice summation techniques (e.g., Ewald summation), but instead define the simulation system to consist 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, space group symmetry is rigorously enforced.

RESULTS AND DISCUSSION

Selection of Test Data Sets

For this work, a total of 33 globular proteins from the PDB are chosen. We chose our test cases based on the following criteria:

1. All proteins are globular X-ray crystal structures with <2.2 Å resolution.
2. The probed helix consists of at least 8 residues (i.e., at least 2 full helix turns). This is to ensure that (a) the helix axis is well defined and (b) the energy contribution of the helix itself dominates the whole loop–helix–loop region, so that it is possible to get an accurate helix position even in the absence of accurate flanking loop conformations. Investigation of methods to predict short helices will be presented in future publications.
3. The helix is not significantly "bent" or "kinked."
4. There are no "obvious problems" in the native protein structure (i.e., large numbers of disordered residues, severe steric clashes, etc.).

Criteria 3 and 4 are admittedly somewhat vague. We have made no effort to quantify the bending of helices, and our selection of test cases with regard to criterion 3 is made primarily based on visual inspection. Future work will investigate the effects of bends and kinks on prediction accuracy; at that point, we will attempt to quantify these features.

Table I lists all the test cases, along with any relevant information associated with them, including the residue ranges associated with the helices and flanking loops. It should be noted that we have chosen to predict a maximum of 8 residues in each flanking loop; that is, if the flanking loops exceed 8 residues, we consider any residues beyond the first 8 to be part of the fixed protein "body." We believe this to be a reasonable choice for most homology modeling applications, because it permits significant helix movements (typically up to at least 5 Å), but the loop prediction component of the algorithm remains computationally tractable. We indicate the chain ID to which the probed helices belong; however, the simulations include all chains in the PDB file, and take crystal packing effects into consideration,[50] as discussed in the Methods section. The last column of Table I indicates whether there is any interaction with ligands or ions in the probed loop–helix–loop region, which can have a significant effect on the prediction results. For targets involved in such interactions, Table II provides more specific information about the type of ligand and interaction distances.

Other criteria were also employed in choosing this test set, anticipating future work that will entail testing the helix prediction algorithm in the context of homology modeling. Specifically, all test cases have homologs in the PDB, in which the helix corresponding to the ones tested here are in a significantly different position relative to the remainder of the protein; that is, in a homology modeling context, the probed helix would contribute significantly to the overall backbone RMSD prior to refinement. We also ensure that the probed helix has few direct contacts with other helices in the same protein, thus eliminating the need to model correlated helix movements.

Effect of Screens

In Table III, we list the number of helix conformation before and after various screens have been applied (see Methods section, step 1, for details), along with the CPU time for screening. [All CPU times refer to 1.4 GHz Pentium III processors.] With only modest CPU time (usually less than 10 min, with a few exceptions), the screens significantly reduce the size of the sampling space by enormous factors. The helix length constraint reduces the astronomical numbers of all possible pairs of grid points ($\sim 10^6$–10^8) down to a tractable range ($\sim 10^3$–10^4). The subsequent additional screens further reduce these numbers to a range that is acceptable for all-atom modeling analysis ($\sim 10^2$).

Initial Helix Sampling

In a vast majority of cases, the initial helix sampling stage (see Methods section, steps 1–3) is able to identify a reasonable starting conformation for the helix (Fig. 3); that is, the lowest energy structure has a reasonably accurate helix position (RMSD <1.0 Å), although the flanking loops may not be accurate at this stage, because no energy-based optimization has been performed on them. The ability to identify accurate helix positions among the many conformations generated (typically up to 6–8 Å RMSD from the native helix position; see Table IV) demonstrates the effectiveness of our energy function and validates our hypothesis that the energy will be dominated by the position of the helix, as opposed to the details of the loops, which are not optimized at this stage.

Moreover, the side-chain optimization and energy minimization frequently improve the accuracy of the helix conformations. As can be seen in Table IV, the average final helix backbone RMSD is 0.9 Å, which is even lower than the average "minimum" RMSD (1.2 Å) (i.e., of the lowest RMSD helix generated in the initial sampling). The average RMSD of the lowest energy helices prior to the side-chain optimization and minimization is 1.6 Å. These results are unsurprising if the native helix position indeed represents a deep well in the potential energy surface,

TABLE I. The Test Set Employed to Validate the Helix Prediction Algorithm

PDB	Chain	reslo	Helix_lo	Helix_hi	reshi	E(native)	Het
1aba		7	15	26	30	−4287.30	N
1ads		129	137	149	155	−13664.31	N
1akh	A	92	97	107	110	−5974.49	N
1ar5	A	53	58	80	88	−18379.68	Y
1b0u		102	110	120	125	−11946.06	N
1b1c	A	115	123	134	137	−8634.66	N
1bvq	A	16	24	42	48	−5733.83	N
1cpq		29	33	46	51	−6100.45	N
1cqd	A	62	70	80	88	−38101.21	Y
1d4a	A	103	110	119	127	−46948.40	Y
1dyr		96	98	108	115	−8879.68	N
1e3w	A	63	68	82	87	−9931.55	Y
1frb		129	137	149	152	−14363.71	N
1fxd		34	41	49	54	−2995.00	N
1im8	A	62	68	75	81	−10407.53	Y
1jap	A	98	106	121	126	−6927.93	N
1jd1	A	41	49	66	73	−31627.03	N
1jse		17	25	36	42	−5768.03	Y
1mka	A	71	79	96	102	−7968.61	N
1mn2		76	84	96	101	−14285.42	N
1qd9	A	38	46	63	67	−5599.83	Y
1qr2	A	103	110	119	127	−9858.06	N
1qu9	A	38	46	63	70	−5872.32	N
1ra9		17	25	35	41	−6961.02	N
1rcd		38	45	72	80	−8961.14	N
1sbc		125	133	144	147	−7558.26	N
1st3		94	102	114	118	−10452.51	N
1tpf	A	12	18	29	37	−10279.41	N
1udg	A	157	165	179	183	−9616.08	Y
1ypc	I	28	32	42	46	−3042.18	N
1ypi	A	11	17	28	35	−8463.96	N
3mds	B	65	70	89	97	−17859.45	Y
3nul		36	44	55	63	−6038.75	N

For each target, we provide PDB code, chain identifier (if applicable), total number of residues in the protein (Nres), starting and ending residue numbers of the whole loop–helix–loop region (reslo and reshi), and starting and ending residue numbers of the helix body (helix_lo and helix_hi). E(native) represents the energy of the native conformation after side-chain optimization and minimization, and is provided for comparison with energies of predicted conformations. The "Het" column indicates whether the probed region interacts with ligands or metal ions (see Table II).

TABLE II. Detailed Information About Interactions Between Loop–Helix–Loop Regions and Heteroatom Groups

		Minimum Distance to		
PDB	Ligand	N-term loop	Helix	C-term loop
1ar5	Fe		2.14	
1cqd	THJ	4.64		
1d4a	FAD	2.79	3.79	5.74
1e3w	SO4	2.81	2.91	
1mka	DAC	3.42	3.61	
1qd9	ACY		4.25	
1qr2	FAD	1.84	3.48	5.92
1tpf	DMS	4.23		
1vfr	FMN	5.37	4.52	
3mds	MN3		2.08	

Only those targets possessing such interactions in the probed regions are listed. Interactions are identified based on a distance cutoff of 4.5 Å for nonmetal ligands and 6.5 Å for metal ions.

allowing the side-chain optimization and energy minimization to drive nearby conformations toward this energy minimum.

Table V presents detailed results for the helix sampling stage, including RMSDs for the helix body, the two flanking loops, and the total loop–helix–loop region; the absolute all-atom energy and relative energy change; and the total CPU time used for the helix sampling stage. There are two cases for which the helix sampling fails to generate a reasonable helix position as the lowest energy prediction at this stage: 1rcd (1.9 Å) and 3nul (2.3 Å). These two failures can be attributed to inadequate sampling, rather than an inaccurate energy function, because there is a positive energy gap between the predicted and native helix conformations. For example, for 1rcd, the predicted structure has $dE = 52.3$ kcal/mol, and in fact, no low RMSD conformations are ever generated; the best helix conformation generated has an RMSD of 3.1 Å. We expect that finer resolution might help in such cases. To confirm this

TABLE III. Summary of Results for Screening of Candidate Helix Positions

Target	N_tot	N_len	N_final	Time (min)
1aba	815,616	35856	779	1.8
1ads	6,640,128	2952	123	4.0
1akh	4,633,220	2496	98	1.4
1ar5	57,874,050	7000	280	22.2
1b0u	1,336,608	16,512	474	1.3
1b1c	785,088	18384	633	1.2
1bvq	6,253,296	5328	222	2.3
1cpq	2,670,624	12,792	494	0.6
1cqd	25,239,120	3096	129	51.5
1d4a	25,067,232	2208	92	31.3
1dyr	2,680,992	11,280	469	0.4
1e3w	6,640,008	11,928	492	4.3
1frb	5,871,600	1584	66	1.5
1fxd	861,120	18,000	412	0.2
1im8	830,304	14,928	416	0.6
1jap	3,687,912	8976	360	4.1
1jd1	6,602,112	6624	276	13.8
1jse	6,564,600	1992	83	2.4
1mka	141,426,000	912	38	68.5
1mn2	11,877,168	1512	63	4.5
1qd9	141,133,824	3504	146	26.0
1qr2	25,338,096	1704	71	11.8
1qu9	6,602,256	1872	78	1.8
1ra9	2,196,144	12,936	476	2.9
1rcd	5,343,408	5976	249	3.4
1sbc	756,960	11,568	421	0.5
1st3	1,928,808	8880	340	0.9
1tpf	3,772,944	9984	358	1.9
1udg	3,542,400	10,080	387	2.2
1ypc	2,094,672	17,280	441	1.1
1ypi	6,462,720	12,960	461	10.2
3mds	58,368,700	3550	142	16.5
3nul	848,232	75,696	1467	2.8

N_tot gives the total potential number of candidates (the product of the number of grid points inside the two bounding spheres for each end of the helix). N_len is the number of candidates that survive the helix length constraint, which is typically in the order of 1% of N_tot. N_final is the number of candidates after all screens are applied. "Time" gives the total CPU time used for screening (in minutes).

TABLE IV. Average Values of Helix RMSDs at Various Stages of the Prediction Algorithm

	Avg. RMSD (Å)
Maximum RMSD after screens	7.32
Minimum RMSD after screens	1.18
RMSD before side-chain optimization	1.61
RMSD after side-chain optimization	0.89

The first two rows refer to the minimum and maximum helix RMSD among those surviving the screening process (generally hundreds of candidates). The last two rows refer to the RMSD of the lowest energy helix before and after side-chain optimization and minimization.

hypothesis, we performed more extensive sampling on these two cases, this time using an initial grid resolution of 0.75 Å. The results, shown in Table VI, demonstrate significant improvement in both the all-atom energy and the helix RMSD. The RMSDs for the lowest energy conformations after the initial sampling stage are now 0.7 Å for 1rcd and 1.6 Å for 3nul.

Another interesting case is the target 3mds. This target demonstrates a particularly strong tendency to preserve the roll angle observed in the native protein. When the along-axis rotation ("roll") deviation with respect to the native position is set to the default value of 7.5°, there is no accepted conformation after the screening stage. When this parameter is reduced to 2.5°, 142 conformations are retained (out of a total of 58,368,700) at the end of the helix sampling, none of which has a roll angle other than 2.5° relative to the native position. In real homology modeling applications, such a phenomenon would undoubtedly provide us with a strong "signal" of the correct "roll" degree of freedom.

Effect of Crystal Packing

Crystal packing[50] plays an important role in prediction accuracy for only 2 test cases. Table VII compares the helix sampling results with and without consideration of crystal packing for two targets: 1akh and 1cqd. The consideration of crystal packing in these two cases clearly makes a significant difference in the sampling results. For 1cqd, consideration of crystal packing effect reduces the RMSD of the whole loop–helix–loop region to 0.8 Å, in contrast to the case where it is ignored (4.5 Å); the differences are primarily in the flanking loop regions. For 1akh, crystal packing affects the accuracy of the helix body itself: 2.7 Å RMSD without crystal packing and 0.5 Å with crystal packing. Nevertheless, in the remaining cases there is either little or no difference in the final results whether considering crystal packing or not (detailed data not shown).

Refinement of Flanking Loops

The results of flanking loop refinement are summarized in Table VIII. For each target, the 5 helix conformations with the lowest all-atom energies at the end of the helix sampling stage are selected for loop refinement. A detailed table with prediction results, including all refined conformations for each target, is presented in Supplementary Materials. As discussed above, the loop refinement algorithm, which has been published previously,[25] is applied to the 2 flanking loops independently and in parallel. The results are combined, and side-chain optimization and complete energy minimization are applied to the entire loop–helix–loop region. This refinement protocol typically leads to significant reductions in the total energy and, in many cases, improved RMSDs in the loop regions (Fig. 4). The effects on average/median loop RMSDs are summarized in Table IX. The improvement obtained in overall loop–helix–loop RMSD is quite substantial. The average RMSD decreases from 1.65 to 1.25 Å, while the median RMSD decreases from 1.00 Å to 0.82 Å. Most of the improvement results from improved loop conformations, particularly in N-terminal loops. [The source of the asymmetry in N- and C-terminal loop accuracy is not clear, and may just be statistical noise.] There are some modest improvements in helix position as well.

Summary of Results

The results clearly validate our hypothesis that the helix–protein interaction dominates the energetics that

TABLE V. Summary of the Results of Helix Sampling

Target	Energy	dE	RMSD(N)	RMSD(H)	RMSD(C)	RMSD(T)	CPU time (min)
1aba	−4280.12	7.18	3.92	0.64	0.43	2.45	954
1ads	−13666.42	−2.10	0.56	0.58	0.25	0.51	520
1akh	−5958.96	15.53	0.49	0.48	0.24	0.44	227
1ar5	−18373.29	6.39	0.76	0.42	1.38	0.80	3318
1b0u	−11886.24	59.82	2.40	1.44	4.09	2.63	547
1b1c	−8620.22	14.44	7.77	0.92	0.18	4.88	344
1bvq	−5734.20	−0.37	8.39	1.05	0.67	4.45	341
1cpq	−6115.79	−15.34	0.88	0.71	0.72	0.74	178
1cqd	−38077.50	23.71	0.50	0.89	0.94	0.76	3614
1d4a	−46890.24	58.16	0.21	1.29	6.52	3.95	2543
1dyr	−8872.04	7.64	0.30	0.32	0.57	0.43	199
1e3w	−9905.72	25.83	0.48	0.65	0.48	0.58	442
1frb	−14349.11	14.60	1.25	0.42	0.38	0.84	561
1fxd	−2988.85	6.15	0.45	0.53	1.16	0.72	248
1im8	−10379.96	27.58	1.15	1.19	1.70	1.37	496
1jap	−6884.58	43.35	2.58	0.62	0.97	1.55	346
1jd1	−31554.96	72.07	1.25	1.18	2.43	1.57	1673
1jse	−5706.43	61.61	7.00	0.48	3.57	4.52	519
1mka	−7943.53	25.09	0.74	0.25	0.39	0.46	1030
1mn2	−14270.91	14.51	3.56	0.54	2.77	2.50	364
1qd9	−5596.61	3.22	1.47	0.57	0.37	0.87	84
1qr2	−6410.35	159.42	3.20	0.63	1.24	1.99	792
1qu9	−5854.51	17.81	1.19	0.27	0.22	0.69	341
1ra9	−6935.41	25.61	5.60	0.92	0.31	3.42	836
1rcd	−8908.80	52.34	7.43	1.91	0.33	3.45	562
1sbc	−7554.82	3.44	1.02	0.57	0.35	0.76	2270
1st3	−10463.00	−10.49	0.37	0.22	0.44	0.32	411
1tpf	−10275.38	4.03	1.76	0.45	0.52	1.00	434
1udg	−9613.35	2.72	2.41	0.62	0.67	1.48	587
1ypc	−3035.35	6.83	0.29	0.35	0.58	0.40	18
1ypi	−15575.35	7.03	1.90	0.81	0.62	1.17	2703
3mds	−17863.81	−4.36	0.66	0.41	0.26	0.42	2224
3nul	−5991.33	47.42	2.81	2.34	2.45	2.48	612

For each target, 4 RMSDs are reported: for the N-terminal loop [RMSD(N)], for the helix alone [RMSD(H)], for the C-terminal loop [RMSD(C)], and for the entire loop–helix–loop region [RMSD(T)]. All RMSD values are in Angstrom units. The final energy and the relative energy change [dE = E(sampling) − E(native)] are also reported. The final column indicates the total CPU time for the helix sampling stage, including helix position enumeration, screening, clustering, loop closure, side-chain optimization, and minimization.

TABLE VI. Comparison of Total Energies (kcal/mol) and Helix RMSDs (Å) with 2 Sampling Resolutions, for the Two "Sampling Error" Cases: 1rcd and 3nul

	Sampling Resolution = 1 Å		Sampling Resolution = 0.75 Å	
Target	Energy	Helix RMSD	Energy	Helix RMSD
1rcd	−8908.8	1.91	−8926.8	0.68
3nul	−5991.3	2.34	−5995.0	1.62

TABLE VII. Effect of Crystal Packing on the Helix Sampling Results

	Without Crystal Packing				With Crystal Packing			
	RMSD(N)	RMSD(H)	RMSD(C)	RMSD(T)	RMSD(N)	RMSD(H)	RMSD(C)	RMSD(T)
1akh	1.18	2.73	1.86	2.18	0.49	0.48	0.24	0.44
1eqd	1.92	0.67	7.48	4.48	0.50	0.89	0.94	0.76

Two targets (1akh, 1eqd) are included. The RMSDs are defined identically to those in Table VI: H = Helix, N = N-terminal flanking loop, C = C-terminal flanking loop, and T = total loop–helix–loop region.

determine the conformation of the loop–helix–loop region. Scoring of the helix alone (assuming sufficiently dense sampling) successfully locates accurate positions for the helix (i.e., to within ~1 Å in a large majority of cases). The success of this strategy is critical in making the problem tractable, because extensive loop sampling over many

TABLE VIII. Results of Flanking Loop Refinement

Target	E(refine)	dE	RMSD(N)	RMSD(H)	RMSD(C)	RMSD(T)
1aba	−4287.70	−0.40	3.84	0.66	2.56	2.63
1ads	−13,684.60	−20.29	0.36	0.59	0.57	0.51
1akh	−5984.50	−10.01	0.77	0.63	0.41	0.64
1ar5	−18,399.21	−19.53	0.95	0.46	1.02	0.69
1b0u	−11,942.97	3.09	0.41	0.80	2.73	1.43
1b1c	−8646.00	−11.34	6.34	0.65	0.19	3.98
1bvq	−5752.05	−18.22	1.72	0.86	0.50	1.09
1cpq	−6135.87	−35.42	1.13	0.77	0.62	0.80
1cqd	−38,120.50	−19.29	0.28	0.40	0.61	0.45
1d4a	−46923.98	24.42	0.24	0.45	4.75	2.87
1dyr	−8883.24	−3.56	0.38	0.49	0.95	0.69
1e3w	−9941.02	−9.47	0.31	0.45	0.53	0.42
1frb	−14,386.20	−22.49	0.54	0.50	0.31	0.50
1fxd	−3028.89	−33.89	1.23	0.79	1.76	1.27
1im8	−10,413.13	−5.60	1.29	0.86	1.92	1.45
1jap	−6934.27	−6.34	0.88	0.33	0.53	0.58
1jd1	−31,619.30	7.73	0.98	0.82	0.31	0.79
1jse	−5736.43	31.60	6.15	0.58	4.36	4.27
1mka	−7980.25	−11.64	0.42	0.29	0.40	0.35
1mn2	−14,286.65	−1.23	2.62	1.03	1.56	1.81
1qd9	−5616.06	−16.23	1.45	0.24	0.34	0.82
1qr2	−9825.52	32.54	1.22	0.43	2.30	1.56
1qu9	−5888.54	−16.22	0.26	0.17	0.21	0.21
1ra9	−6969.67	−8.65	3.08	1.08	0.27	1.99
1rcd	−8971.25	−10.11	1.27	0.43	1.54	0.94
1sbc	−7569.31	−11.05	0.85	0.59	0.33	0.68
1st3	−10,472.39	−19.88	0.43	0.25	0.30	0.32
1tpf	−10,298.09	−18.68	1.74	0.39	0.76	1.03
1udg	−9646.37	−30.29	4.38	0.73	1.60	2.65
1ypc	−3047.33	−5.15	0.29	0.27	0.89	0.51
1ypi	−8510.58	−46.62	1.83	0.86	0.65	1.16
3mds	−17,901.34	−41.89	0.27	0.49	0.87	0.60
3nul	−6038.57	0.18	1.67	1.71	1.74	1.65

For each target, the resulting RMSD for the N-terminal loop [RMSD(N)], RMSD for the helix body [RMSD(H)], RMSD for the C-terminal loop [RMSD(C)], RMSD for the whole loop–helix–loop region [RMSD(T)], as well as the final energy [E(refine)] and the relative energy change [dE = E(refine) − E(native)] are reported. All reported energy and RMSD data are those after full side-chain optimization and energy minimization of the whole loop–helix–loop region.

TABLE IX. Average/Median RMSDs (Defined as in Table VI) Before and After Loop Refinement

	Before Loop Refinement	After Loop Refinement
RMSD(N)	2.27/1.25	1.50/0.98
RMSD(H)	0.75/0.62	0.60/0.58
RMSD(C)	1.16/0.58	1.16/0.65
RMSD(T)	1.65/1.00	1.25/0.82

helix positions would be extremely expensive. We should note however that there might be cases where the present approach breaks down, such as very short helices (less than 2 turns) not investigated here.

The accuracy of the final protocol is extremely encouraging for an initial effort. The average helix RMSD is only 0.6 Å, and the worst case is only 1.7 Å. Further improvement of the helix position can likely be achieved by additional refinement stages. For the present, the results are, to our knowledge, substantially more accurate and robust than can be achieved using simple empirical scoring functions to position the helix, and approach the precision needed for realistic biological applications, including structure-based drug design. The key test with regard to this application is docking of active ligands into a homology model where a helix shift is critical to correct construction of the binding pocket in the target; we intend to pursue such assessments in the near future.

The flanking loop accuracy is generally quite satisfactory. In a significant fraction of cases where there are nontrivial errors in the loop, similar errors are also present in prediction of the loop using the native helix position (e.g., 1aba, 1mn2, 1udg). The errors in the native loop prediction (presented in Table X) show a similar distribution to the more extensive tests of the loop prediction methodology described previously.[25] In some cases, the errors may be related to close interactions of the loop with ligands or metals (1d4a, 1qr2), which are not modeled in the present calculations. It is interesting to note that in cases (1ar5, 1qd9) where the helix, but not the flanking loops, interacts closely with a ligand or metal, there is no obvious degradation in accuracy relative to other cases

TABLE X. Summary of the Loop Prediction Results on the Native Protein Structures

Target	E	dE	RMSD(N)	RMSD(C)
1aba	−4287.91	−0.61	3.03	0.22
1ads	−13,644.89	19.42	0.22	0.25
1akh	−6010.25	−35.76	0.30	0.17
1ar5	−18,460.30	−80.62	0.21	1.20
1b0u	−11,938.55	7.51	0.25	0.20
1blc	−8614.89	19.77	0.86	0.11
1bvq	−5705.46	28.37	3.31	0.21
1cpq	−6086.50	13.95	0.20	0.97
1cqd	−38,800.32	−699.11	0.38	0.27
1d4a	−47,275.96	−327.56	0.22	4.30
1dyr	−8859.98	19.70	0.08	1.94
1e3w	−9920.51	11.04	0.14	0.10
1frb	−14,345.50	18.21	0.25	0.13
1fxd	−2922.91	72.09	0.97	1.35
1im8	−10,426.54	−19.01	1.05	0.32
1jap	−6910.15	17.78	4.74	0.49
1jd1	−31,880.28	−253.25	0.44	0.18
1jse	−5758.89	9.14	0.41	0.17
1mka	−7976.90	−8.29	0.35	0.48
1mn2	−14,283.49	1.93	0.36	0.37
1qd9	−5619.94	−20.11	1.24	0.28
1qr2	−9856.57	1.49	4.65	0.36
1qu9	−5872.81	−0.49	0.28	0.15
1ra9	−6946.10	14.92	0.41	0.65
1rcd	−8947.57	13.57	0.28	0.30
1sbc	−7519.62	38.64	1.37	0.21
1st3	−10,456.25	−3.74	2.38	0.19
1tpf	−10,273.07	6.34	1.36	0.36
1udg	−9632.91	−16.83	2.22	0.42
1ypc	−3032.47	9.71	0.28	0.21
1ypi	−8479.66	−15.70	1.93	0.46
3mds	−17,995.61	−136.16	0.29	0.39
3nul	−6030.79	7.96	0.75	0.22
	Average RMSD		1.07	0.53
	Median RMSD		0.41	0.28
	Average RMSD (N- & C-term combined)		0.80	
	Median RMSD (N- & C-term combined)		0.35	

We start from the native conformation, in which there is no helix position deviation, and rebuild the two flanking loops, followed by full side-chain optimization and energy minimization on the whole loop–helix–loop region. For each target, the resulting RMSD for the N-terminal loop [RMSD(N)] and C-terminal loop [RMSD(C)], as well as the final energy and the energy change relative to the native (dE), are reported. Summary statistics, consisting of the average and median RMSDs, are also listed.

lacking ligands. This is consistent with the idea that positioning of the helix is driven by a relatively deep energy minimum, which is created by many residues of the helix simultaneously interacting with the body of the protein. Such a minimum is better able to survive perturbing interactions with a ligand or metal. In contrast, the flexibility of some loops may make them more susceptible to perturbation by a nearby ligand or metal ion.

There are a few cases where the accuracy of flanking loops in the helix prediction results is significantly worse than the accuracy of the loop predictions when the helix is held at the native conformation. These cases include 1b0u, 1blc, 1jse, 3nul, and 1rcd. Of these, 1b0u and 1blc appear to represent failures of the energy model; that is, with the helix allowed to move, the energy function identifies an incorrect conformation as having significantly lower energy than the native conformation. Improvement of these cases may require modifications of the potential energy function and/or solvation model, but may also simply reflect errors in the assignment of protonation states to titratable side-chains such as histidine. We have discussed 3nul and 1rcd previously; finer initial helix sampling improves accuracy in these cases (indeed, it may prove desirable to use the parameters that succeeded in these two cases as the default; alternatively, some way of identifying particularly difficult sampling problems needs to be defined and implemented). The most serious failure of the sampling algorithm is seen in 1jse. Despite the excellent prediction of the helix position (0.6 Å RMSD), the flanking loops both have very large RMSDs and the total energy is quite far above the native energy. This is the only case where one can observe a strong and problematic sensitivity of the flanking loop structure to the helix position on the tenths of Angstrom scale. The obvious

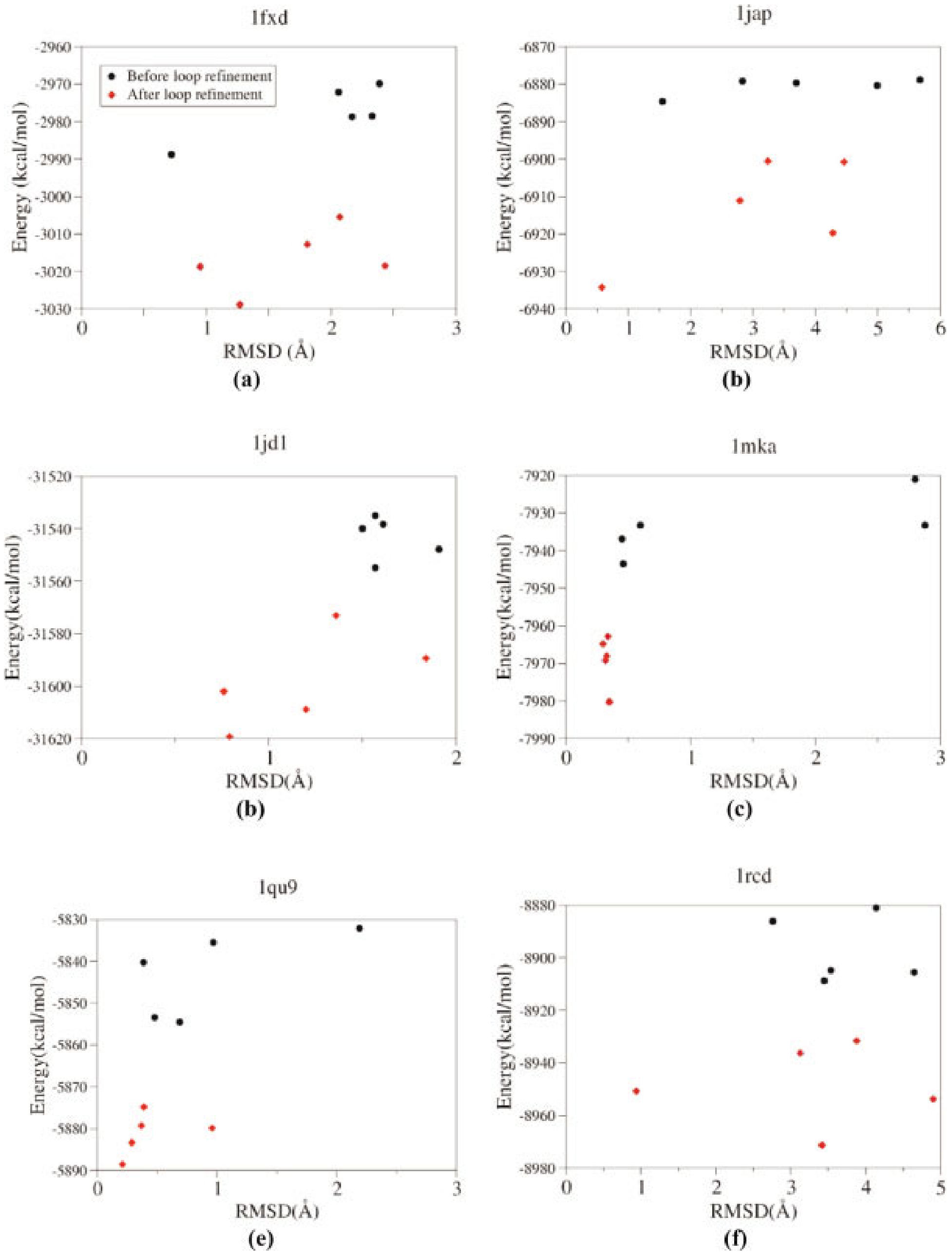

Fig. 4. Plots of energy (kcal/mol) versus RMSD (Å) for the top 5 (or whatever the number of remaining clusters at the end of the helix sampling stage) structures before and after iterative loop rebuilding for selected targets (**a–f**).

strategy to improve this case is to refine the position of the helix and repredict the flanking loops, and this strategy may represent the logical path to obtaining higher precision for the methodology in general.

CONCLUSIONS

We have developed a practical methodology for predicting and refining loop–helix–loop regions of globular proteins with high precision (~1 Å RMSD) in the great majority of test cases that have been examined. The accuracy is attributable to the combination of the physics-based, all-atom energy function (including continuum solvent) and the hierarchical screening and sampling protocol. The predictions typically require tens of hours of CPU time and can be trivially distributed over several processors.

There remain a small number of cases (the most prominent being 1jse) where sampling clearly limits accuracy. Additional rounds of refinement, focusing in a finer grid around the initially located helix position, should effectively deal with problems of this type. In a few cases, the sampling of the loop–helix–loop region generates conformations with significantly lower energies than the minimized native. These errors could indicate limitations of the fixed-charge force field or the implicit solvent model, and we are investigating specific classes of errors associated with charged side-chains, using as well problematic cases encountered in side-chain and loop prediction. However, the "energy errors" may also reflect two other limitations of our methodology. First, we fix all protonation states of the titratable side-chains at the predominant states for neutral pH (Asp/Glu negatively charged, Lys/Arg positively charged, His protonated at the delta position). This method does not take into account shifts in pKa's of the side-chains due to the protein environment, and errors in protonation state could manifest as "energy errors." Second, the contributions of internal protein entropy to relative free energies are largely ignored in the current methodology; only side-chain entropy is treated even approximately (see Methods section). This limitation should be most apparent when multiple low energy conformations exist, and/or the helix or surrounding loops are not well packed against the remainder of the protein. We have no strong evidence of errors arising from neglecting entropic effects, although at the present time, it is admittedly difficult to identify unambiguously the causes of the various energy errors.

The most important remaining challenges involve integrating the algorithm into a strategy for homology model refinement. An early version of the helix prediction algorithm was employed in the CASP5 competition, with some evident success (significantly improved helix conformations in targets T132, T133, T150, and T178).[24] However, numerous challenges remain. First, in homology modeling applications, the helix prediction algorithm must be tolerant of errors in the structure of the surrounding regions of the protein, or provide a means for optimizing both simultaneously. In this work, we have performed validation tests that are analogous to the tests reported in majority of papers on loop prediction (i.e., the remainder of the protein is held rigidly fixed at the native conformation). In realistic applications, incorrect conformations of the side-chains and/or loops in the vicinity of the sampled loop–helix–loop region could strongly impact the quality of the prediction, if they are held rigidly fixed. We are currently developing an approach in which side-chains surrounding the loop–helix–loop region are removed during the initial sampling stage and optimized simultaneously with the loop–helix–loop side-chains during the refinement stages. Preliminary, anecdotal results—including the CASP5 results mentioned above, which used an early version of this algorithm—suggest that this strategy is tolerant of moderate errors in the surrounding protein and can retain high accuracy for realistic homology modeling cases. Validation and tuning of the modified algorithm on a large, diverse test set is under way and will be reported in a future publication.

Second, in some homology modeling applications, correlated interactions among multiple helices could be important; that is, simultaneous refinement of multiple helices may be necessary (e.g., for all-α proteins, as opposed to the refinement of single helices discussed here). Third, corresponding helices among homologs can differ not only in position/orientation but also in their precise length. The differences in length among homologous helices are usually not large, typically 1 or 2 resides (i.e., 1/2 helix turn or less), but these differences can contribute significantly to backbone RMSD. Finally, nonconservative amino acid substitutions in the middle of helices, especially to Gly or Pro, can also lead to helix kinking (i.e., a breakdown in our assumption of a rigid helix). Changes in helix size can be addressed via our loop prediction algorithm (extended into the ends of the helical regions), but the coupling of this type of calculation with helix shift prediction needs to be investigated in some detail.

ACKNOWLEDGMENTS

We thank David L. Pincus (Columbia U. Department of Chemistry) and Tyler Day (Schrödinger, Inc.) for testing early versions of the code. X. Li would like to thank David L. Pincus and Susan Harrington for many stimulating discussions and suggestions.

REFERENCES

1. Barthe P, Roumestand C, Demene H, Chiche L. Helix motion in protein C12A-p8(MTCP1): comparison of molecular dynamics simulations and multifield NMR relaxation data. J Comput Chem 2002;23:1577–1586.
2. Elber R, Karplus M. Multiple conformational states of proteins—a mulecular-dynamics analysis of myoglobin. Science 1987;235:318–321.
3. Rojewska D, Elber R. Molecular-dynamics study of secondary structure motions in proteins—application to myohemerythrin. Proteins 1990;7:265–279.
4. Furois-Corbin S, Smith JC, Kneller GR. Picosecond timescale rigid-helix and side-chain motions in deoxymyoglobin. Proteins 1993;16:141–154.
5. Bright JN, Shrivastava IH, Cordes FS, Sansom MSP. Conformational dynamics of helix S6 from Shaker potassium channel: simulation studies. Biopolymers 2002;64:303–313.
6. Adamian L, Liang J. Helix–helix packing and interfacial pairwise interactions of residues in membrane proteins. J Mol Biol 2001;311: 891–907.
7. Tuffery P, Etchebest C, Popot JL, Lavery R. Prediction of the positioning of the 7 transmembrane alpha-helices of bacteriorhodopsin–a molecular simulation study. J Mol Biol 1994;236:1105–1122.
8. Mingarro I, Eloffsson A, vonHeijne G. Helix–helix packing in a membrane-like environment. J Mol Biol 1997;272:633–641.
9. Lear JD, Gratkowski H, Adamian L, Liang J, DeGrado WF. Position-dependence of stabilizing polar interactions of asparagine in transmembrane helical bundles. Biochemistry 2003;42: 6400–6407.
10. Fain B, Levitt M. A novel method for sampling alpha-helical protein backbones. J Mol Biol 2001;305:191–201.
11. Nanias M, Chinchio M, Pillardy J, Ripoll DR, Scheraga HA. Packing helices in proteins by global optimization of a potential energy function. Proc Natl Acad Sci USA 2003;100:1706–1710.
12. Huang ES, Samudrala R, Ponder JW. Distance geometry generates native-like folds for small helical proteins using the consensus distances of predicted protein structures. Protein Sci 1998;7: 1998–2003.
13. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
14. Mumenthaler C, Braun W. Predicting the helix packing of globular proteins by self-correcting distance geometry. Protein Sci 1995;4:863–871.
15. Cohen FE, Richmond TJ, Richards FM. Protein folding—evaluation of some simple rules for the assembly of helices into tertiary structures with myoglobin as an example. J Mol Biol 1979;132:275–288.
16. Kamiya Y, Reynolds CA. Brownian dynamics simulations of the

beta(2)-adrenergic receptor extracellular loops: evidence for helix movement in ligand binding? J Mol Struct Theochem 1999;469: 229–232.

17. Falke JJ, Hazelbauer GL. Transmembrane signaling in bacterial chemoreceptors. Trends Biochem Sci 2001;26:257–265.
18. Brzozowski AM, Pike ACW, Dauter Z, Hubbard RE, Bonn T, Engstrom O, Ohman L, Greene GL, Gustafsson JA, Carlquist M. Molecular basis of agonism and antagonism in the oestrogen receptor. Nature 1997;389:753–758.
19. Chothia C, Lesk AM. Helix movements in progeins. Trends Biochem Sci 1985;10:116–118.
20. Chothia C, Lesk AM. Helix movements and the reconstruction of the heme pocket during the evolution of the cytochrome-C family. J Mol Biol 1985;182:151–158.
21. Chothia C, Lesk AM, Dodson GG, Hodgkin DC. Transmission of conformational change in insulin. Nature 1983;302:500–505.
22. Gerstein M, Lesk AM, Chothia C. Structural mechanisms for domain movements in proteins. Biochemistry 1994;33:6739–6749.
23. Tramontano A, Leplae R, Morea V. Analysis and assessment of comparative modeling predictions in CASP4. Proteins 2001;S5:22–38.
24. 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. 2003. Submitted for publication.
25. Jacobson MP, Pincus DL, Rapp CS, Honig B, Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins 2003. Forthcoming.
26. Fiser A, Do RKG, Sali A. Modeling of loops in protein structures. Protein Sci 2000;9:1753–1773.
27. Bruccoleri RE, Karplus M. Chain closure with bond angle variations. Macromolecules 1985;18:2767–2773.
28. Bruccoleri RE, Karplus M. Prediction of the folding of short-polypeptide segments by uniform conformational sampling. Biopolymers 1987;26:137–168.
29. Das B, Meirovitch H. Optimization of solvation models for predicting the structure of surface loops in proteins. Proteins 2001;43:303–314.
30. deBakker 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;51:41–55.
31. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 2000;40:135–144.
32. 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.
33. Fidelis K, Stern PS, Bacon D, Moult J. Comparison of systematic search and database methods for constructing segments of protein-structure. Protein Eng 1994;7:953–960.
34. Moult J, James MNG. An algorithm which predicts the conformation of short lengths of chain in proteins. J Mol Graph 1986;4:180.
35. Rapp CS, Friesner RA. Prediction of loop geometries using a generalized Born model of solvation effects. Proteins 1999;35:173–183.
36. 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.
37. 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.
38. Smith KC, Honig B. Evaluation of the conformational free-energies of loops in proteins. Proteins 1994;18:119–132.
39. Zhang H, Lai L, Wang L, Han Y, Tang Y. A fast and efficient program for modeling protein loops. Biopolymers 1997;41:61–72.
40. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: The colony energy and its application of the problem of loop prediction. Proc Natl Acad Sci USA 2002;99:7432–7437.
41. van Vlijmen HWT, Karplus M. PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
42. 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.
43. 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.
44. Jorgensen WL, Maxwell DS, Tirado-Rives 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.
45. Donate LE, Rufino SD, Canard LHJ, Blundell TL. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: a database for modeling and prediction. Protein Sci 1996;5:2600–2616.
46. Hartigan JA. Clustering algorithms. New York: Wiley; 1975.
47. Hartigan JA, Wong MA. Algorithm AS 136: a K-means clustering algorithm. Appl Stat 1979;28:100–108.
48. Xiang ZX, Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 2001;311:421–430.
49. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. J Phys Chem B 2002;106:11673–11680.
50. Jacobson MP, Friesner RA, Xiang ZX, Honig B. On the role of the crystal environment in determining progein side-chain conformations. J Mol Biol 2002;320:597–608.
51. Christopher JA, Swanson R, Baldwin TO. Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods. Comput Chem 1996;20:339–345.
52. Xie DX, Schlick T. Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications. Siam J Optimiz 1999;10:132–154.
53. Schlick T, Overton M. A powerful truncated Newton method for potential-energy minimization. J Comput Chem 1987;8:1025–1039.
54. Schlick T, Fogelson A. Algorithm 702—Tnpack—a truncated Newton minimization package for large-scale problems; 1. Algorithm and usage. ACM Trans Math Software 1992;18:141–141.
55. Tuckerman M, Berne BJ, Martyna GJ. Reversible multiple time scale molecular dynamics. J Chem Phys 1992;97:1990–2001.
56. Ghosh A, Rapp CS, Friesner RA. Generalized Born model based on a surface integral formulation. J Phys Chem B 1998;102:10983–10990.