PROTEINS: Structure, Function, and Genetics 35:173–183 (1999)

Prediction of Loop Geometries Using a Generalized Born Model of Solvation Effects

Chaya Sendrovic Rapp[1] and Richard A. Friesner[1,2]*
[1]*Department of Chemistry and Center for Biomolecular Simulation, Columbia University, New York, New York*
[2]*Department of Biochemistry and Molecular Biophysics and Center for Biomolecular Simulation, Columbia University, New York, New York*

ABSTRACT **We have carried out an extensive exploration of the possibility of predicting the structure of long loops in proteins, using an 8- and a 12-residue loop in ribonuclease A as models. The native X-ray structure is used as a template while allowing for template flexibility; this makes our work relevant to the problem of homology modeling in which the template is not precisely known. Energies are calculated with the AMBER* and AMBER94 molecular mechanics potentials and the generalized Born continuum solvation model; and conformational space is sampled by means of a combination of Monte Carlo and molecular dynamics methods. Our AMBER94 results demonstrate that we can successfully generate loops with low root-mean-square deviations from the native as well as excellent energetic rankings. Proteins 1999;35:173–183.**

Key words: molecular dynamics; Monte Carlo; protein folding; simulated annealing; tertiary structure

INTRODUCTION

The determination of loop geometries in proteins is a central problem in biomolecular modeling. From a theoretical viewpoint, the ability to predict loop structures in the Protein Data Bank (PDB) provides a demanding, but potentially tractable, test of protein potential energy functions, solvation models, and sampling algorithms. At the same time, loop prediction is one of the most difficult aspects of the construction of homology models, which is becoming an increasingly important method of generating atomic resolution protein structures without the necessity of an experimental X-ray crystallographic or NMR structure determination.[1]

Over the past decade, there have been a number of efforts to develop algorithms for predicting loop structures. Most of these algorithms can be characterized by the following general structure:

1. A set of initial guess loops are generated, either by random selection (satisfying endpoint constraints),[2–4] by a construction procedure,[5] or by selecting loops from the PDB.[6,7]
2. The loop is inserted into the protein, and refinement of the structure is carried out via Monte Carlo, simulated annealing, minimization, or some combination thereof.[8–11] To enhance sampling, the hard-core van der Waals term is often modified,[12] and techniques such as multiple copy simulations[13] can be used to reduce computation time. Some approaches have a number of stages in the simulations, with the final one utilizing the most accurate potential model.[4]
3. The final energies obtained from the full set of independent runs are collected, and the correlation of RMS deviation with the target loop is ascertained.

The quality of the results obtained from such calculations is dependent on two factors: 1) accuracy of the potential functions used to rank the candidate loops, and 2) efficacy of the sampling process. In the final refinement stage, the potential model that is usually employed is a conventional protein molecular mechanics force field; to date, the CHARMm[14] force field has been most frequently used. While the quantitative precision of CHARMm and other molecular mechanics models is not entirely clear, there is no question that simple packing and dispersive interactions are described quite reasonably. On the other hand, the treatment of solvent in the great majority of calculations has been highly approximate; methods include distance-dependent dielectric, complete elimination of electrostatic charges, and simplified continuum treatments based on surface area.[15] Alternatively, solvent effects are included only in the final phase of energy, evaluation rather than during the course of a simulation, thus generating unrealistic initial conformations.[16,17]

Most of the results that have been reported in the papers cited above have been for relatively short loops, typically with a length of seven residues or fewer. Furthermore, the remainder of the protein is ordinarily held rigid, and endpoint constraints are rigorously enforced. These conditions ensure that the residues near the ends of the loop are almost certain to prefer the native structure based on trivial structural considerations; there is some flexibility

Grant sponsor: National Science Foundation Graduate Research Fellowships, Grant sponsor: National Institutes of Health Institute of General Medicine; Grant number: GM-52018.
Dr. Rapp is currently at the Department of Chemistry, Yeshiva University, New York, New York.
*Correspondence to: Dr. Richard A. Friesner, Department of Chemistry, Columbia University, 3000 Broadway, New York, NY 10027.
E-mail: rich@chem.columbia.edu
Received 4 May 1998; Accepted 10 December 1998

of the residues in the middle, but the rigid native template provides a strong selection of the native loop based on packing arguments. These considerations help to explain the low backbone RMSDs that have been reported in many of these experiments despite the deficiencies in the potential function.

There are certainly a class of practical problems where calculations along these lines would prove quite valuable. However, there are also obvious limitations. First, for longer loops, there is significant conformational flexibility available to the loop even if the template is perfectly rigid and known to high accuracy. In these cases, there are likely to be a significant number of dissimilar multiple minima, which the potential function must be able to discriminate. The ability of models with crude treatment of solvation to do this is suspect on physical grounds, and has not yet been demonstrated in practice. Second, in a realistic homology modeling context, the template structure is known only approximately. In this situation, it seems highly likely that allowing the region of the protein around the loop to relax will be essential to achieving reliable predictions. If the side chains of the loop in the target protein are significantly different from those of the homologous protein that is being used as a template, one would expect the template to be significantly reorganized in the actual structure of the target; surely, the ability to pick out the native loop on the basis of shape complementarity alone will then be substantially compromised. As very few of the methods cited above have been tested in an actual homology modeling application, it is difficult to know just how problematic this issue is in practice; however, the lack of reports of successful building of long loops in homology modeling efforts suggests that the difficulties are considerable.

In this article, we directly confront the two issues discussed above. First, we employ, along with a molecular mechanics protein potential, an approximate solvation model—the generalized Born (GB) model of Still and coworkers[18]—which we and others have shown provides remarkably good agreement with accurate Poisson-Boltzmann calculations for the relative solvation free energies of peptide and protein conformations.[19] The GB model, as implemented in the Macromodel[20] molecular modeling code has an analytical gradient and is sufficiently fast to allow lengthy minimization and simulated annealing runs of substantial fractions of a protein. This, in conjunction with the commitment of substantial computer resources, allows us to tackle the prediction of long loop segments while allowing template flexibility. Although not as demanding as a true homology modeling simulation, these calculations represent a significant upgrading of the level of realism (and hence of difficulty, as well as compared to the work discussed previously). Second, we initiate an effort to examine the effects of the details of the protein potential function and continuum solvation model on the reliability of the results. Currently, both components of the methodology contain both random and systematic errors, which will be reduced as better models are developed. However, it is important to obtain an assessment of the range of performance of currently available approaches. We have chosen to study two models: (1) AMBER*/GB as implemented in Macromodel and (2) AMBER94/GB as implemented in Macromodel. The performance of these two potentials is contrasted for the two examples presented below.

We have chosen to study two loops (residues 13–24 and residues 64–71) from the protein ribonuclease A, for which there are high resolution X-ray crystallographic and nuclear magnetic resonance (NMR) determinations of structure. The backbone root-mean-square deviation (RMSD) between the X-ray (1rat.pdb) and NMR (2aas.pdb) PDB structures is 1.11 Å; the two structures are pictured in Figure 1. Both of the loops we studied have virtually identical conformations in the X-ray and NMR structures; this indicates that the loops assume single, well-defined structures in solution as well as in the crystal, and that the accuracy of the structures (loops are often determined rather imprecisely by NMR) is unambiguously high. Therefore, a correct energy model and search algorithm should locate these conformations as global minima (assuming they are not metastable states—this is just the usual hypothesis adopted in protein structure prediction). As both loops exhibit a number of distinctive structural motifs which do not appear to be trivially determined by packing alone, this represents a major computational challenge.

The article is organized as follows: in the following section, we describe the computational model and the simulations algorithms employed to generate structural predictions. Then, results are presented for the two loops discussed above. Finally, in the conclusion, we discuss future directions of this research.

COMPUTATIONAL MODEL AND SIMULATION METHODS

Protein Potential Function

As discussed above, we investigate here two protein potential functions, both molecular mechanics type potentials in which atomic nuclei are treated as classical point particles moving on an approximate Born-Oppenhiemer potential energy surface. The first, AMBER*, is the implementation of the original AMBER potential of Kollman and coworkers[21] in Macromodel.[20] AMBER* is therefore part of the older generation of protein molecular mechanics force fields, which one would expect to exhibit performance that is not as accurate as that of some of the newer models. AMBER* employs a united atom representation in which all atoms are included explicitly with the exception of hydrogens bonded to carbon. The second potential function is the all atom AMBER94 potential, the most recent version of AMBER,[22] again, we employ the implementation in Macromodel. For both potential functions, we use infinite cutoffs for all nonbonded calculations.

Generalized Born Solvation Model

We use a generalized Born/surface area (GB/SA)[18] approximation for the protein–solvent and solvent–solvent interactions. This is a relatively inexpensive continuum solvent model and is thus ideal for simulations on a large

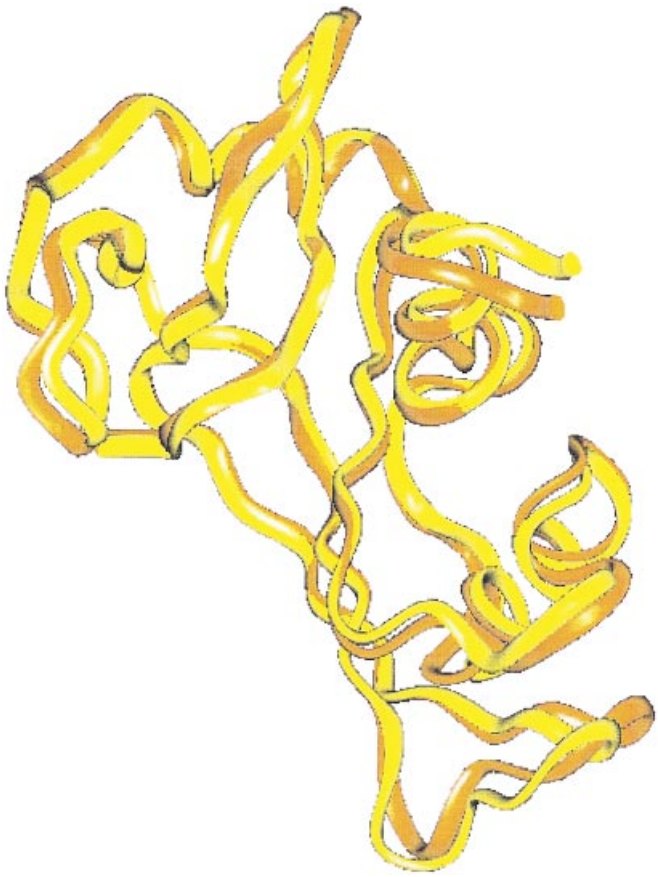

Fig. 1. Ribonuclease A: Native X-ray (yellow) and NMR (orange) structures superimposed; backbone RMSD = 1.11 Å.

system containing a sizable protein in water. The free energy of solvation in this model is written as the sum of three terms, a solvent–solvent cavity term, a solvent–solute van der Waals term and a solvent–solute polarization term. The sum of the first two terms is directly proportional to the total solvent accessible surface area, and the third term is approximated by a modified generalized Born equation.

In a previous paper,[19] we demonstrated the superiority of the generalized Born model in evaluating solvation free energies for molecular systems. This was done based on a comparison of a number of approximate solvation models with an accurate solution of the Poisson-Boltzmann equation for a large dataset of peptide structures, ranging from a single amino acid to a sequence of nine residues. The generalized Born model provided reasonably good solvation free energies and was capable of reproducing the Poisson-Boltzmann energetic rankings of different peptide conformations rather well. In contrast, distance-dependent dielectric and surface area-based methods failed badly. Thus we consider the generalized Born model to be the most accurate solvation model that can at present be efficiently used within the context of simulating a large and flexible biological system.

Fig. 2. Residues 64–71: Loops of native X-ray (yellow) and NMR structures (orange) when the two structures are completely superimposed; backbone RMSD = 1.68 Å.

Minimization and Simulated Annealing

Atomic level refinement is accomplished through a combination of minimization and molecular dynamics simulated annealing. In molecular dynamics the classical equations of motion for the nuclei are solved numerically to give the system's trajectory through phase space. The SHAKE[23] algorithm is used to constrain the bonds to hydrogen and to allow for a timestep of 1.5 fs. Simulated annealing is an optimization scheme that attempts to circumvent entrapment in local minima by allowing for uphill moves that bring the system over energy barriers into new low-energy wells.[24] The simulation is begun at a high temperature, which is reduced according to a cooling schedule. At a given temperature, the system moves from one conformation to another with a probability proportional to a Boltzmann factor. Conjugate gradient minimization is used to direct the system to the bottom of a particular energy well.

To locate the optimal configuration for the loop under consideration it is necessary to screen a large number of initial structures in a cost-efficient way. For this purpose, we use substructure calculations in which simulations are performed on only the loop and its neighboring region. In the Macromodel/Batchmin program, the atoms in the substructure are either completely flexible or constrained to move in place around a narrow harmonic well; the constrained atoms serve as a boundary region around the region of interest. All interactions are calculated within the flexible region as well as nonbonded interactions with the boundary region, while within the boundary region, only bond stretching interactions are calculated. Atoms not included in the substructure are excluded from the calculation. In the context of this article, the term "minimization" refers to complete structure minimization, while "substructure minimization" or "substructure annealing" are used to refer to the above-described method.

Clustering

To understand how the generated loop structures are related to one another, we employed X-cluster, the clustering program in Macromodel. As input to X-cluster, we use a group of structures, each positioned such that it is superimposed on the appropriate native structure. The program then searches for clusters using the RMSD of the loop backbone atoms as the distance criterion. X-cluster works by means of a hierarchical scheme in which clusters at high clustering levels are unions of clusters that form at low levels. The user chooses an appropriate clustering level based on experiment.

Simulation Protocol

Generation of initial loops

After a great deal of experimentation, we developed the following protocol for generating candidate loop structures. We utilize the X-ray structure minimized in water (using either the AMBER*/GB or AMBER94/GB model) as a template for the protein without the target loop. We refer to this template as the "minimized native." A set of 130 alternate loop conformations are produced by means of a Monte Carlo algorithm using a reduced protein model (in which side chains are represented by a Cβ only) in a modified version of Gunn's TRIP program.[25] In this program, the loop is represented by a set of backbone dihedral angles ϕ and ψ, associated with each residue along the chain. The relative orientation of the loop between the two secondary structure elements at its ends is specified by including with the severed loop two Cα secondary structure atoms on each end and calculating six independent geometric parameters, which constitute the endpoint geometry of the loop. The input to the program is a set of randomly generated strings of dihedral states that serve as the starting loop geometries. New loops are generated by means of trial moves in which three residue loop segments are replaced with values from a table. The endpoint geometry is used as the target function of the simulation. For each generated loop an error function is used to evaluate how much the endpoint geometry of the new loop deviates from that of the native. If the deviation is less than a cutoff on the order of 10^{-4} Å, the new loop is saved.

Once a suitable backbone structure is obtained, an all-atom model is generated by adding explicit side chains with the RSA program of Shenkin and coworkers.[26] This is a fast, automated side-chain prediction algorithm in which a Monte Carlo algorithm is used to sample rotamer states from a rotamer library and a simulated annealing scheme is used to minimize collisions between side chains. The program produces four distinct results, representing four initial guesses for the side chains of the loop; this ensures that the phase space of the side chains is explored in addition to that of the backbone. For our purposes, it is only necessary to achieve a reasonable guess for the side chains, as, subsequently, they will be refined in the context of the complete protein. Following side chain addition, the severed loop is reinserted into the protein by a superposition of the Cα secondary structure atoms at the ends of the generated loop on the corresponding atoms of the original structure[27] to produce 520 complete atomic level structures. Pairwise, RMSDs are then calculated for all the generated loops to ensure that they are dissimilar (RMSD $\geq$ 1 Å) and thus cover a wide region of phase space.

Loop refinement via Macromodel

The complete atomic level structures are then subject to a three-stage refinement protocol in the Macromodel/Batchmin molecular modeling program using the AMBER* or AMBER94 potential function, and a generalized Born/surface area (GB/SA) approximation for the solvent.

Stage 1: A substructure minimization (up to 10,000 iterations) is performed on the loop plus a surrounding region of 3 Å, with an additional 2 Å included in the calculation as a boundary region. Energies are evaluated for the substructure region of each minimized structure.

Stage 2: The 30 (or 50) lowest-energy structures produced by stage 1 are then subject to molecular dynamics substructure annealing according to the following schedule: 10 ps molecular dynamics equilibration at 300 K followed by a linear cooling of the bath temperature from 300 K to 50 K over the next 40 ps. Annealing is followed by complete minimization and evaluation of total energies.

Stage 3: A group of structures are selected from stage 2 to undergo stage 3, a long run of simulated annealing involving 50 ps molecular dynamics equilibration at 300 K (or 500 K) followed by a linear cooling of the bath temperature to 50 K over the next 500 ps.

The structures selected to undergo stage 3 are the lowest-energy structures that represent each unique loop backbone appearing among the 10 lowest-energy structures produced by stage 2 (i.e., if 2, 3, or 4 structures that are side-chain variants of a common backbone appear among the 10 lowest-energy structures, only the lowest-energy structure in the set is subject to stage 3). The long annealing run is followed by complete minimization and evaluation of total energies.

The objective of this hierarchical sampling protocol is to obtain both broad coverage of phase space and accurate evaluation of the relative energetics of competitive structures. Stage 1 allows sampling of a large, diverse set of initial loops, and selects from the initial list those loops in basins of attraction with reasonable energies. Stage 2 provides a significantly better evaluation of the attainable energies of those basins, allowing more structures to be discarded. Stage 3 represents our best estimation of these energetics, with the large cost of each individual annealing calculation made tractable by the substantial reduction in the number of starting geometries over stages 1 and 2.

Definition of RMSD

RMSDs refer to angstroms of the backbone atoms of the loop of a particular structure from the loop of the appropriate minimized native, when the structure and the mini-

mized native are completely superimposed. This is a critical distinction from papers in the literature in which the RMSD is defined by loop superposition alone, a much less demanding criterion. It is obviously possible for the loop to be poorly oriented with respect to the protein, which would produce a qualitatively incorrect structural prediction, while still yielding a good "RMSD" according to the criterion of bare loop superposition.

All loops that are illustrated are shown as they appear within the complete structure from which they have been severed when that structure is completely superimposed on the appropriate native template.

RESULTS

Overview

For each of the two loops investigated below, the protocols of the preceding section were applied using both the AMBER*/GB and AMBER94/GB models. We present results for both potentials, side by side, at the first two stages of the sampling process. However, for the AMBER*/GB model, the poor results obtained at stage 2 discouraged the investment of further efforts using that model, and thus stage 3 was performed only for AMBER94.

Additionally, structures obtained from the native loop were generated and tracked at each stage for comparison. These simulations based on the native structure provide energetic benchmarks, which allow the validity of the energetic model to be evaluated; if the energy of a nonnativelike structure obtained in a simulation is far below the best energies generated by simulated annealing starting from the native structure, this suggests that the energetic model is seriously in error. On the other hand, if the entire set of energies generated in the sampling process are significantly above the energies produced during simulations starting from the native, this is indicative of a failure of the sampling algorithm to locate the native basin of attraction.

Ribonuclease A; Residues 64–71

The first loop we chose to study, residues 64–71 in the protein ribonuclease A, is eight residues in length and is pictured in Figure 2. It is situated between two beta sheets at the leftmost part of the protein, which is pictured in its entirety in Figure 1. For this loop, the X-ray and NMR structures are slightly different in conformation (RMSD = 1.68 Å), but essentially represent a common topology that should be possible to identify for this medium-size loop.

For each minimized native template, a set of 130 initial loop conformations were generated by TRIP and built up to 520 complete structures with pairwise loop RMSDs ranging from 1 to 8 Å. Stage 1, substructure minimization, was performed using the both the AMBER* and AMBER94 potential functions. The following setup was used for the substructure: 20% of the structure's atoms, comprising the loop and a surrounding region of 3 Å, are included in the completely flexible region and and additional 9%, comprising a boundary region of two angstroms are included in the boundary region. Figures 3 and 4 are plots of the substructure energies (the energy of the flexible region) of the minimized structures and their loop RMSDs to the appropriate minimized native for the two potentials. Each plot includes data for 350–400 structures; the remaining structures contain high-energy loops and were eliminated early in the calculation (after approximately 100 iterations). Both plots show that among the 30 lowest-energy generated structures, there is at least one that has a low RMSD (RMSD $\sim$ 1.2 Å) to the minimized native.

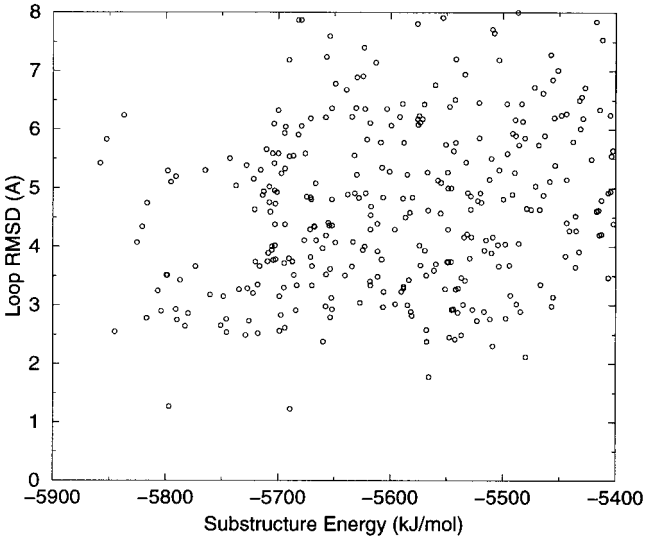

Fig. 3. Residues 64–71; AMBER*: substructure energies and loop RMSD of 363 structures after substructure minimization screen.

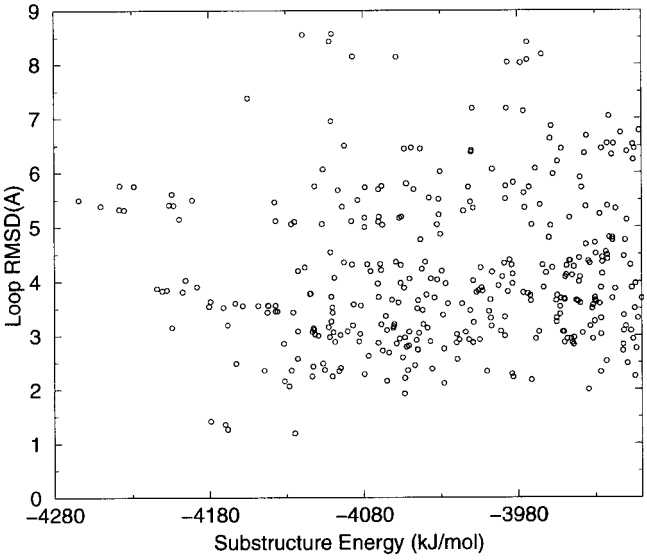

Fig. 4. Residues 64–71; AMBER94: substructure energies and loop RMSD of 390 structures after substructure minimization screen.

Following the initial substructure minimization screen, the 30 lowest energy structures in each set were subject to stage 2, substructure annealing (using the same setup as for substructure minimization), and complete minimization. Figures 5 and 6 are plots of the total energies and loop

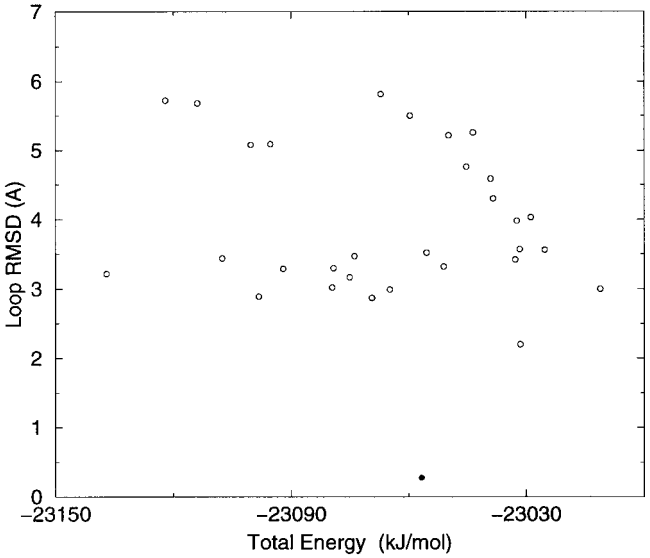

Fig. 5. Residues 64–71; AMBER*: energies and loop RMSD of 30 structures after annealing and complete minimization; ''native'' structure indicated with a filled symbol.

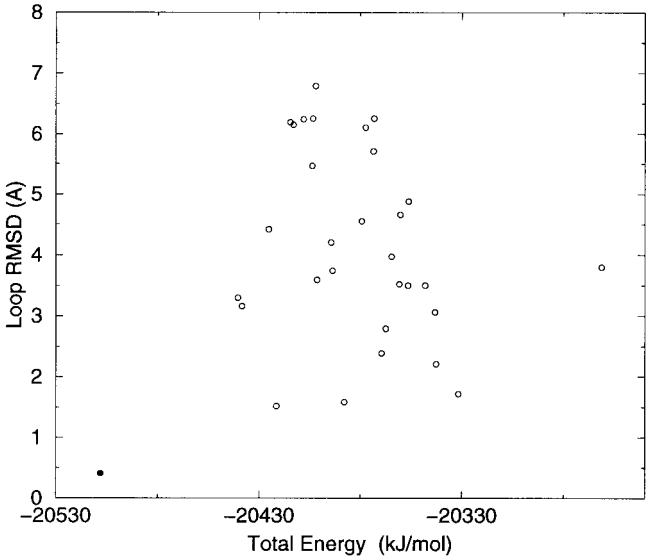

Fig. 6. Residues 64–71; AMBER94: energies and loop RMSD of 30 structures after annealing and complete minimization; ''native'' structure indicated with a filled symbol.

RMSDs of the completely minimized structures for each potential. The structures that result when the same protocol of annealing and minimization is performed on the minimized native are indicated with filled symbols. For the structures produced by the AMBER* model, the annealed native has a very high energy compared to the others; this is a surprising result and is probably indicative of a serious qualitative error in the potential function. By contrast, the AMBER94 potential ranks this structure at an energy ~70 kJ/mol below the lowest-energy-generated structure (RMSD = 3.30 Å). The lowest RMSD structure generated by AMBER94 (RMSD = 1.52 Å) is ranked as the fourth lowest-energy structure.

The 30 lowest-energy structures generated using AMBER94 were clustered in X-cluster using a clustering level of 15 (i.e., a division into 16 clusters). Data is reported for all clusters with three or more members. The three clusters that appear have 6, 4, and 4 members and are pictured in Figures 7–9. The lowest RMSD structure is included within the third cluster indicating that this conformation represents a basin of attraction that is repeatedly being located by the conformational search.

Stage 3 was applied to seven of the structures produced by the AMBER94 model in stage 2, including the annealed native. Each structure was subject to a lengthy simulated annealing run starting from a temperature of 500K, and then completely minimized. Table I, in which X refers to the annealed native, compares the total energies of these seven structures after stage 2 and stage 3, and lists their final loop RMSDs. One notes that with the exception of the annealed native, which is already a highly optimized structure, all the other structures undergo a decrease in energy of at least 50 kJ/mol after annealing in stage 3. Most significantly, the lowest RMSD (RMSD = 1.46 Å) structure, labeled structure 20, is now ranked as the lowest-energy structure with an energy that is within 6 kJ/mol of the annealed native. This structure is pictured in Figure 10 and is almost identical in geometry to the native.

Ribonuclease A; Residues 13–24

The second loop we chose to study, residues 13–24, is situated between two alpha helices at the exterior of the right side of the protein and is pictured alone in Figure 11. The RMSD between the loop in the X-ray and NMR structures is 1.14 Å, indicating an almost identical conformation.

For each minimized native template, a set of 130 initial loop conformations were generated by TRIP and built up to 520 complete structures with pairwise loop RMSDs ranging from 1 to 7 Å. Stage 1, substructure minimization, was performed using both the AMBER* and AMBER94 potential functions. The following setup was used for the substructure: 23% of the structure's atoms, comprising the loop and a surrounding region of 3 Å, are included in the completely flexible region and and additional 19%, comprising a boundary region of 2 Å are included in the boundary region. Figures 12 and 13 are plots of the substructure energies (the energy of the flexible region) of the minimized structures and their loop RMSDs to the appropriate minimized native for the two potentials. Each plot includes data for 350–400 structures; the remaining structures contain high-energy loops and were eliminated early in the calculation (after approximately 100 iterations). Both plots show that among the 30 lowest-energy-generated structures, there is at least one that represents a reasonable starting point for refinement (RMSD ~ 2.5 Å).

Following the initial substructure minimization screen, the 50 lowest-energy AMBER* minimized structures and the 30 lowest-energy AMBER94 minimized structures were subject to stage 2, substructure annealing (using the

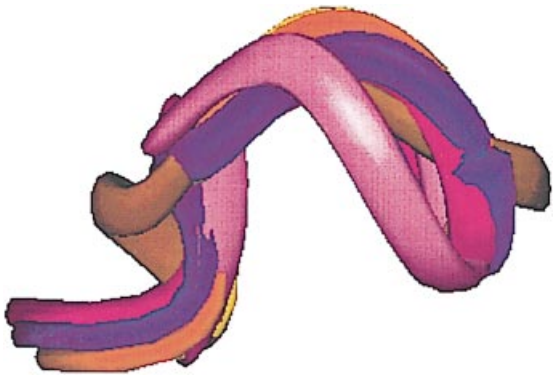

Fig. 7. Residues 64–71; AMBER94: substructure annealing and minimization followed by complete minimization. Loops of first cluster: 4 members.

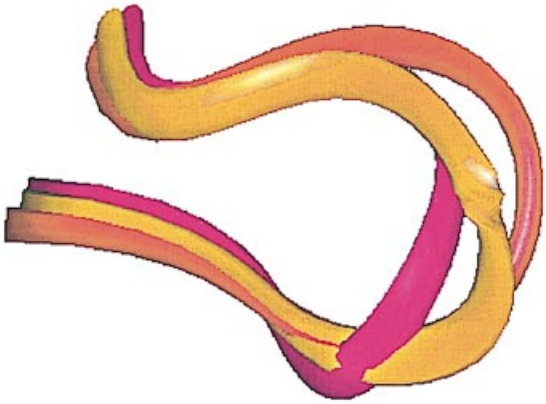

Fig. 8. Residues 64–71; AMBER94: substructure annealing and minimization followed by complete minimization. Loops of second cluster: 4 members.

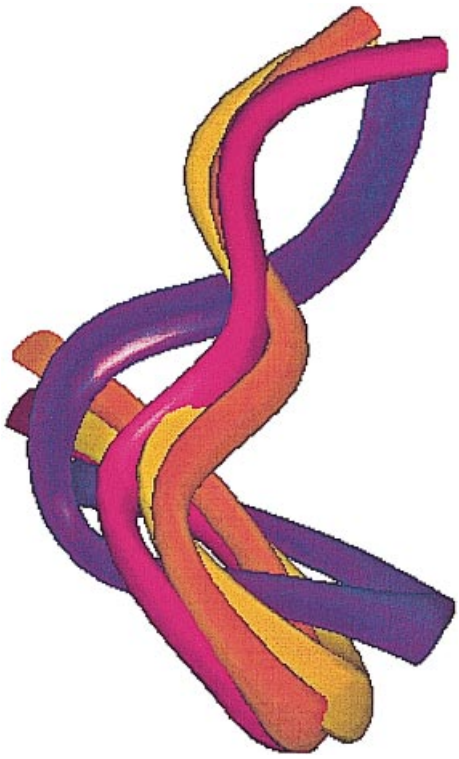

Fig. 9. Residues 64–71; AMBER94: substructure annealing and minimization followed by complete minimization. Loops of third cluster: 4 members.

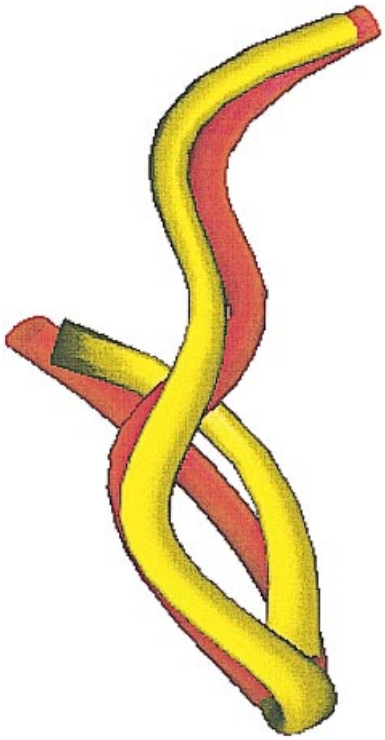

Fig. 10. Residues 64–71; AMBER94: stage 3 annealing and complete minimization; loops of lowest energy structure, structure 20 (red), and minimized native (yellow).

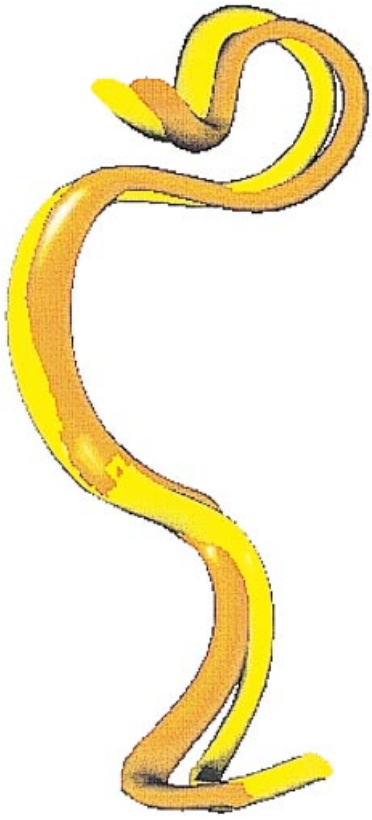

Fig. 11. Residues 13–24: Loops of native X-ray (yellow) and NMR structures (orange) when the two structures are completely superimposed; backbone RMSD = 1.14 Å.

TABLE I. Residues 64–71: Final Set of Structures Produced by Lengthy Annealing and Minimization; Stage 2 and Stage 3 (Final) Energies and Final RMSD

ID	Stage 2 energy	Final energy	Final loop RMSD (Å)
X	−20508.219	−20508.117	1.12
20	−20421.438	−20502.344	1.46
19	−20440.605	−20491.607	3.00
14	−20425.230	−20479.037	4.27
2	−20414.832	−20461.727	5.74
29	−20402.105	−20457.387	6.87

same setup as for substructure minimization) and complete minimization. Figures 14 and 15 are plots of the total energies and loop RMSDs of the completely minimized structures for the two potentials. The structures that result when the same protocol of annealing and minimization is performed on the minimized native are indicated with filled symbols. For the structures produced using AMBER*, the lowest RMSD structure produced by this stage has an RMSD of 2.24 Å and is relatively high in energy. The lowest-energy structure has and RMSD of 4.19 Å and is competitive in energy with the annealed native. For the structures produced using AMBER94, the lowest RMSD structure produced by this stage has an RMSD of 0.80 Å and is ranked as the second lowest-energy structure. The lowest energy structure has and RMSD of 2.50 Å and is approximately 10 kJ/mol above the annealed native in energy.

The 30 lowest-energy structures generated using AMBER94 were clustered in X-cluster using a clustering level of 12 (i.e., a division into 19 clusters). Data is reported for all clusters with four or more members. The two clusters that appear each have four members and are pictured in Figures 16 and 17. The three lowest-energy structures, including the lowest RMSD structure, are included within the first cluster, indicating that this conformation represents a favored minimum.

Stage 3 was applied to seven of the structures produced by the AMBER94 model in stage 2, including the annealed native. Each structure was subject to a lengthy simulated annealing run starting from a temperature of 300 K, and then completely minimized. (A starting temperature of 500 K brought all of the structures to higher energies). Table II, in which X refers to the annealed native, compares the total energies of these seven structures after stage 2 and stage 3 and lists their final loop RMSDs. Four of the seven structures increase in energy, indicating an escape from the original basin of attraction into a higher basin; for these structures we use the original structures resulting from stage 2 for comparison. The other three structures decrease or remain approximately the same in energy. Examining the results of stage 2 and 3 yields the structure labeled 2 (RMSD = 2.50 Å), and illustrated in Figure 18, as the lowest-energy structure. The second lowest-energy structure, labeled 19 (RMSD = 0.80 Å), is shown in Figure 19 and is virtually identical to the native structure.

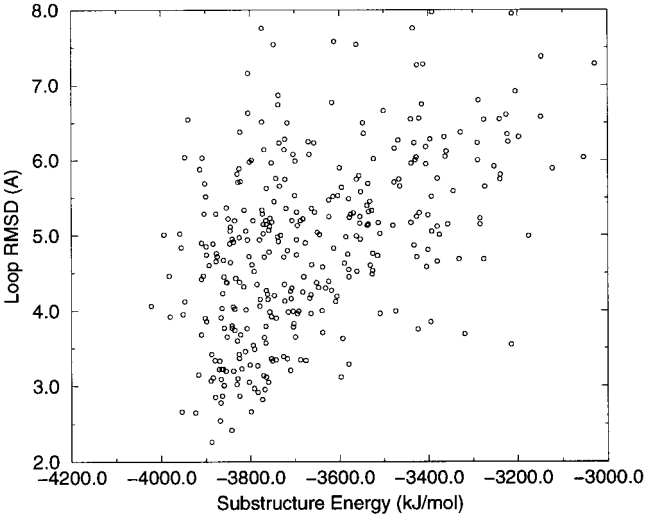

Fig. 12. Residues 13–24; AMBER*: substructure energies and loop RMSD of 363 structures after substructure minimization screen.

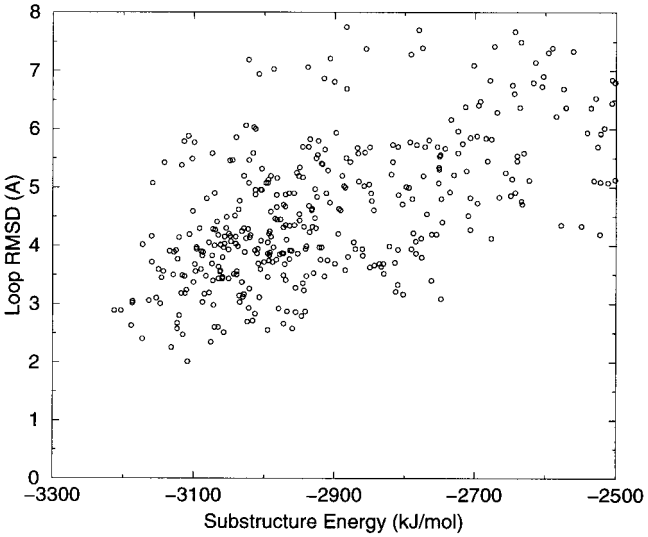

Fig. 13. Residues 13–24; AMBER94: substructure energies and loop RMSD of 390 structures after substructure minimization screen.

Discussion

Several significant conclusions can be drawn from the above results. First, the AMBER* protein potential function is inadequate to describe the relative energetics of loop geometries in proteins. For the eight-residue loop, the energy of a number of misfolded structures were substantially below the best energies obtained from simulated annealing of the native. The performance of the model in generating and ranking loops with good resemblance to the native structure was also qualitatively inferior to that of AMBER94. Considering that the AMBER* potential was developed almost a decade ago, these deficiencies are not entirely unexpected.

By contrast, the performance of the AMBER94 model was quite good, although not perfect; it appears as though

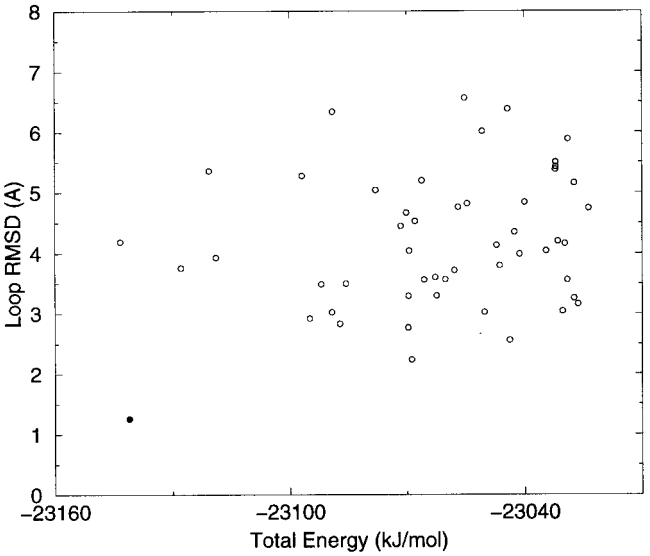

Fig. 14. Residues 13–24; AMBER*: energies and loop RMSD of 50 structures after annealing and complete minimization; "native" structure indicated with a filled symbol.

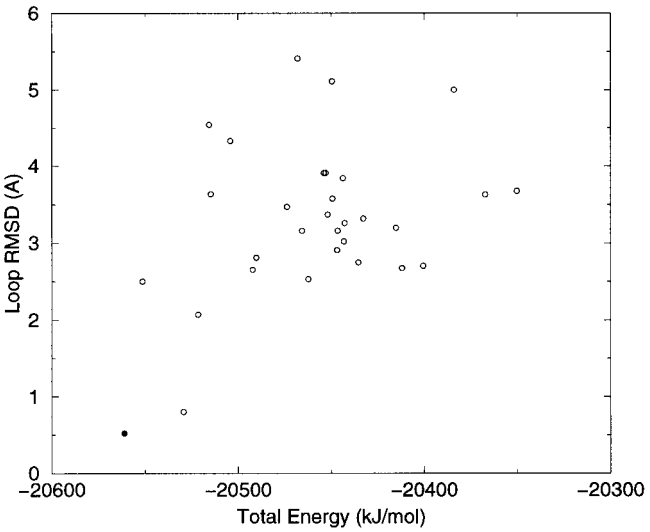

Fig. 15. Residues 13–24; AMBER94: energies and loop RMSD of 30 structures after annealing and complete minimization; "native" structure indicated with a filled symbol.

the second generation of AMBER development has resulted in very substantial improvements in the quality of the force field, demonstrated outside the scope of the data used to parametrize it. Energies generated from the native-based simulations were in both cases lowest in energy, and the model was able to consistently locate structures with remarkably low RMSD from the native in a relatively modest amount of computation time. The fact that the best structures found ranked first (in the case of the 8-residue loop) and second (for the 12-residue loop) is highly encouraging; it suggests that only limited improvements in the sampling algorithm and/or energy model will be required to reliably predict loop geometries, at least under the conditions studied here. There are obvious directions in which to pursue improvements in both of these components, as we briefly discuss in the conclusion.

TABLE II. Residues 13–24: Final Set of Structures Produced by Lengthy Annealing and Minimization; Stage 2 and Stage 3 (Final) Energies and Final RMSD

ID	Stage 2 energy	Final energy	Final loop RMSD (Å)
X	−20561.344	−20559.932	0.49
19	−20529.586	−20544.195	0.81
2	−20551.762	−20538.662	2.50
13	−20514.918	−20531.572	3.65
9	−20515.818	−20499.598	4.55
22	−20473.850	−20473.363	3.53
11	−20468.062	−20461.438	5.40

CONCLUSION

We have carried out the first extensive exploration of the possibility of predicting the structure of long loops in proteins using a molecular mechanics potential function and the generalized Born model of Still and coworkers, which we believe to be the most accurate continuum solvation model currently available that is tractable for sampling conformational space. Using the AMBER94 protein potential function, success in generating loops with low RMSDs from the native and excellent energetic rankings was demonstrated. Furthermore, the length of the loops studied, and the conditions under which the calculations were performed (flexible template structure) rendered the problems posed here among the most challenging investigated in loop prediction studies to date.

As indicated above, improvements in this approach will be pursued in two areas. First, the effects of more extensive sampling will be investigated. This includes generation of additional structures via Monte Carlo, lengthening of the simulated annealing runs, experimentation with the simulated annealing protocol, and investigation of other, quite different sampling algorithms. The results obtained here, in which very low RMS structures were generated for long loops with relatively small computational effort, suggest that the sampling problem is quite tractable, and additional effort, combined with the ever-increasing power of computer hardware, should allow this aspect of the problem to be handled in a convincing fashion. At some point, it will be possible to attribute any remaining problems to the potential functions and solvation model rather than inadequacy of sampling.

Second, we intend to explore the use of different potential functions and solvation models. It is clear from the present article that the protein potential function accuracy is critical in loop prediction. We will first investigate the quality of results obtained with other current generation fixed charged force fields, such as the OPLS-AA potential of Jorgensen and coworkers,[17] and the MMFF potential of Halgren.[12–16] Subsequently, we intend to carry out experiments with a polarizable force field that we are in the

Fig. 16. Residues 13–24; AMBER94: substructure annealing and minimization followed by complete minimization. Loops of first cluster: 4 members.

Fig. 17. Residues 13–24; AMBER94: substructure annealing and minimization followed by complete minimization. Loops of second cluster: 4 members.

process of developing.[2,26] There are also a number of alternative continuum solvation models that are worthy of investigation. These include a new version of the generalized Born model, based on surface rather than volume integration, that we have recently developed,[1] and accurate numerical solution of the Poisson-Boltzmann equation. Because of the computational expense of the latter approach, it would probably be most profitably employed as a single-point screening calculation, or steepest descent minimization, subsequent to a generalized Born simulated annealing/minimization protocol. Single-point screening using Poisson-Boltzmann methods has already been car-

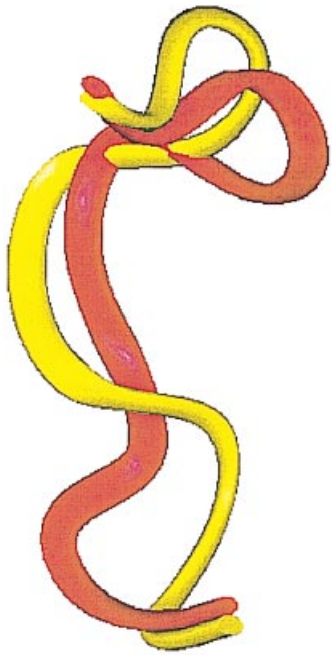

Fig. 18. Residues 13–24; AMBER94: stage 2 annealing and complete minimization; loops of lowest energy structure, structure 2 (red), and minimized native (yellow).

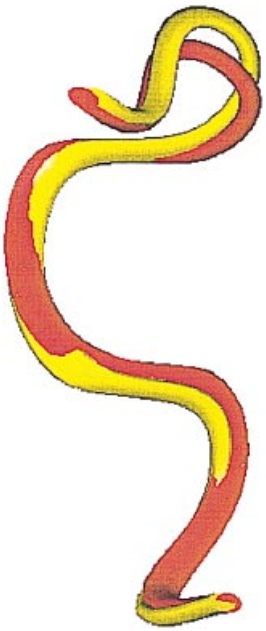

Fig. 19. Residues 13–24; AMBER94: stage 3 annealing and complete minimization; loops of second lowest energy structure, structure 19 (red), and minimized native (yellow).

ried out by Vorobjev and coworkers,[32] and proved successful in discriminating free energies of proposed protein structures.

Finally, the results reported in this article must be regarded as anecdotal, rather than as a statistical demonstration that reliable results for loop prediction can be obtained. Statistical studies in which a large number of loops are investigated are required for the latter objective. As computational power increases and improvements are made in sampling algorithms, this will become a feasible goal.

REFERENCES

1. Aghosh A, Rapp CS, Friesner RA. A generalized born model based on a surface integral formulation. J Phys Chem 1998;102:10983–10990.
2. Banks JL, Kaminski GA, Zhou R, Mainz DT, Berne BJ, Friesner RA. Parameterizing a polarizable force field from ab initio data: the fluctuating point charge model. J Chem Phys 1999;110:741–754.
3. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 1983;4: 187–217.
4. Carlacci L, Englander SW. The loop problem in proteins: a Monte Carlo simulated annealing approach. Biopolymers 1993;33:1271–1286.
5. Carlacci L, Englander SW. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. J Comp Chem 1996;17:1002–1012.
6. Collura V, Greany PJ, Robson B. A method for rapidly assessing and refining simple solvent treatments in molecular modeling: example studies on the antigen-combining loop h2 from fab fragment mcpc603. Protein Eng. 1994;7:221–233.
7. Collura V, Higo J, Garnier J. Modeling of protein loops by simulated annealing. Protein Sci 1993;2:1502–1510.
8. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acid s and organic molecules. J Am Chem Soc 1995;117:5179–5197.
9. 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.
10. Edinger SR, Cortis C, Shenkin PS, Friesner RA. Solvation free energies of peptides: comparison of approximate continuum solvation models with accurate solution of the Poisson-Boltzmann equation. J Chem Phys 1996;101:1190–1197.
11. Gunn JR, Friesner RA. Parallel implementation of a protein structure refinement algorithm. J Comp Chem 1996;17:1217–1228.
12. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. J Comp Chem 1996;17:490–519.
13. Halgren TA. Merck Molecular Force Field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J Comp Chem 1996;17:520–552.
14. Halgren TA. Merck molecular force field. III. Molecular geometrics and vibrational frequencies for MMFF94. J Comp Chem 1996;17:553–586.
15. Halgren TA. Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J Comp Chem 1996;17:616–641.
16. Halgren TA, Nachbar RB. Merck molecular force field. IV. Conformational energies and geometries for MMFF94. J Comp Chem 1996;17:587–615.
17. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the opls all-atom force field on conformational energetics and proper ties of organic liquids. J Am Chem Soc 1996;118:11225–11236.
18. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr 1976;32:922–923.
19. Kirkpatrick Jr S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983;22:671.
20. Mohamadi F, Richards NG, Guida WC, Liskamp R, Lipton M, Caufield C, Chang G, Hendrickson T, Sti II WC. Macromodel-an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. J Comp Chem 1990;11:440.
21. Mosimann S, Meleshko R, James MNG. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 1995;23:301–317.
22. Olivia 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.
23. Pellequer JL, Chen Shu wen W. Does conformational free energy distinguish loop conformations in proteins? Biophys J 1997;73: 2359–2375.
24. Shenkin PS. Prediction and evaluation of side-chain conformations for protein backbone structures. Proteins 1996;26:323–352.
25. Smith KC, Honig B. Evaluation of the conformational free energies of loops in proteins. Proteins 1994;18:119–132.
26. Stern HA, Kaminski GA, Banks JL, Zhou R, Berne BJ, Friesner RA. Parameterizing a polarizable force field from ab initio data: dipole and combined fq-dipole models. J Chem Phys 1999 (in press).
27. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127.
28. Sudarsanam S, DuBose RE, March CJ, Srinivasan S. Modeling protein loops using a phi, psi database. Protein Sci 1995;4:1412–1420.
29. van Gunsteren WF, Berendsen WJC. Algorithms for macromolecular dynamics and constraint dynamics. Mol Phys 1977;34:1311–1327.
30. van Vlijmen HWT, Karplus M. pdb-Based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 1997;267:975–1001.
31. Vasmatzis G, Brower R, DeLisi C. Predicting immunoglobulin-like hypervariable loops. Biopolymers 1994;34:1669–1680.
32. Vorobjev YN, Almagro JC, Hermans J. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 1998;32:339–413.
33. Weiner KJ, Kollman PA, Nguyen DT, Case DA. An all-atom force field for simulations of proteins and nulcleic acids. Phys Rev E 1986;7:230–252.
34. Zhang H, Lai L, Wang L, Han Y, Tang Y. A fast and efficient program for modeling protein loops. Biopolymers 1997;41:61–72.
35. Zheng Q, Kyle DJ. Accuracy and reliability of the scaling-relaxation method for loop closure: an evaluation based on extensive and multiple copy conformational samplings. Proteins 1996; 24:209–217.
36. Zheng Q, Rosenfeld R, Vajda S, DeLisi C. Determining protein loop conformations using scaling relaxation techniques. Protein Sci 1993;2:1242–1248.
37. Zheng Q, Rosenfeld R, Vajda S, DeLisi C. Loop closure via bond scaling and relaxation. J Comp Chem 1993;14:556–565.