**Hongyu Zhang**
**Luhua Lai***
**Leyu Wang**
**Yuzhen Han**
**Youqi Tang**
*Institute of Physical Chemistry*
*Peking University*
*Beijing 100871, China*

# A Fast and Efficient Program for Modeling Protein Loops

*We developed an efficient Monte Carlo Simulated Annealing (MCSA) program for modeling protein loops with high speed. The total conformational energy in each step of MCSA simulation consists of two parts: the nonbonded atomic interaction represented by a simple soft-sphere potential and the harmonic distance constraint to ensure the smooth connection of the loop segment to the rest of the protein structure. The soft-sphere potential was a simplified potential that has been successfully used by the authors in modeling the carbohydrate part of glycoprotein systems [H. Zhang, Y. Yang, L. Lai, and Y. Tang (1996),* Carbohydrate Research, *Vol. 284, pp. 25–34]. It only considers the purely repulsive steric interactions to avoid artificial attractive forces between atoms in the absence of solvent molecules. The N-terminal of the loop segment was connected to the bulk protein part, and two dummy main-chain atoms N and Cα immediately following the C-terminal of the loop segment were constrained to their real positions in the protein structure, which not only assures the correct geometry of loop–protein connection but also is more rigorous than the previous work. To improve the speed, two strategies, the local region method and grid-mapping method, were devised to accelerate the computation of environmental interaction that is responsible for the major part of the computing consumption. The grid-mapping method can reduce computational time dramatically. Conformations with rational steric packing and smooth connection to the rest of the protein structure were generated by the MCSA program, and then were refined by the empirical force field CHARMm [B. R. Brook, R. E. Braccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus (1983),* Journal of Computational Chemistry, *Vol. 4, pp. 187–217*]. Bovine pancreatic trypsin inhibitor *(BPTI) was used as an example to test the ability of loop modeling of the method, and five loops in BPTI were calculated. Conformations close to the crystal structure were generated for all of them. With the criteria of CHARMm energy, near-native conformations can be selected, for example, the backbone rms deviation 0.93 Å from the crystal structure was gotten for the longest 9-residue loop.* © *1997 John Wiley & Sons, Inc.*

## INTRODUCTION

Protein loop modeling is important in protein structural biology for its wide applications. For example, during protein structure prediction, once the core structure is determined from homologous proteins, the surface loop regions, which in most cases are not conserved and includes gaps and chain reversals, need to be modeled afterward. Similarly, in an nmr spectroscopy experiment, due to various reasons, sometimes we may not have enough distance constraints to define all the struc-

Received February 6, 1996; accepted April 11, 1996.

* To whom correspondence should be addressed.

Biopolymers, Vol. 41, 61–72 (1997)

CCC 0006-3525/97/010061-12

ture segments, and part of them have to be generated from other known atomic positions. Another challenge comes from antibody engineering where people expect to alter the nature of the combining site of antibody and its concomitant binding and specificity by experimental mutagenesis on the hypervariable loops.[1] The common point in all the above cases demands that a flexible peptide chain be anchored into a known fixed framework with a proper orientation relative to the bulk protein environment.

Present strategies of loop search can be divided into two main categories[2,3]: knowledge-based approach and ab initio approach. Sometimes, the two are mixed in one program.[4]

## Knowledge-Based Approach

In knowledge-based approach, a conformation found in the loop of equivalent length in the homologous protein, or if unavailable, in another protein, is chosen to be the candidate conformation based on the criteria of anchor deviation, sequence homology, and steric contacts, etc.[5,6] Alternative ways generate the candidate conformations not directly from the protein structure data base, but from the special regulation extracted from the structure data base. The approach developed by Moult et al.[7] involves the selection of a representative set of $\phi$, $\psi$, and $\chi$ values for each residue from the distributions of these angles in refined protein structures, and generating a series of loop fragments by various combinations of these dihedral angles. Those fragments that come close to satisfying the closure requirements are then refined by energy minimization in the presence of the rest of the protein structure. Recently, Sudarsanam et al.[8] created from the Protein Data Bank (PDB) a data base of a list of allowed $\phi_{i+1}$ and $\psi_i$ angles with which they can construct the protein loops. A special method for modeling complementary determining regions (CDRs) of antibody, the Key residue method,[9] showed that there is a small repertoire of main-chain conformations for at least five of the six hypervariable regions of antibodies, and that the particular conformation adopted is determined by a few key conserved residues. Reczko et al.[10] used an artificial neural network trained on a large set of loops from the PDB to predict the most variable antibody hypervariable loop, CDR-H3, with a reasonable success.

Powerful as the knowledge-based method is, it cannot provide reliable results when lacking homologous fragment or the *key* residues. Moreover, some of the methods are specific for a narrow range of loops, e.g., those of antibodies, and are hard to be applied in generic situations.

## Ab Initio Approach

In the ab initio approach, a structure data base is not necessary. Candidate conformations can be created from special algorithms, including Go–Scheraga,[11-15] TWEAK,[16] bond scaling,[17,18] and Monte Carlo.[19-24] Go and Scheraga proposed a method for exactly closed rings in molecules with fixed bond lengths and bond angles.[11] Loops in proteins are special cases of rings in which the two ends of a loop do not coincide. In a ring with $n$ rotatable bonds, there are only $n - 6$ independent ones. The values of the six dependent variables are determined by the conditions of ring closure. Bruccoleri and Karplus[12,13] modified the Go–Scheraga method by allowing bond-angle bending when the equations did not have a solution. In treating loops longer than three residues, they carried out searches over additional dihedral angles. Their flexible-geometry method has been incorporated into the program CONGEN,[13] which was designed to execute loop searches in homology-modeling applications. Dudek and Scheraga[14] and Palmer and Scheraga[15] developed alternative formulations of the equations, involving efficient representation of hydration free energy, a local minimization procedure with respect to subsets of degrees of freedom, and taking into account differences in the backbone geometry of various amino acids. With these methods, they improved computational efficiency, and demonstrated that bond-angle bending is not necessary for chain segments five residues or longer in length. Another approach to loop searches, the TWEAK method developed by Shenkin et al.,[16] is carried out by setting each dihedral angle on the main chain of the variable fragment to a random value, then using an iterated linearized Lagrange multiplier technique to enforce the geometric constraints with minimal conformational perturbation. The random tweak method can avoid the exponential increase in computing time with loop size, which is inherent in systematic searches.

Generally, the above ab initio methods,[11-16] in the first step, generate backbone conformations subject to the imposed distance constraints, followed by a step of screening or minimization to remove those having bad steric contacts and high potential energy within the loop segment or between the loop segment and the environment; then comes the last step of side-chain generation. Their weaknesses are as follows: they have difficulty covering the large conformational space of middle size and long loops (>5), they have a low acceptation ratio in the second step of screening as to reject most of the conformations generated time-consumingly in the first step, and have to model side chains separately that may introduce extra errors to the modeling.

Improved methods consider both the constrained effects of loop terminals and the interaction of the loop with protein environment simultaneously, with side chains generated along with the backbone in the process of simulation. Therefore, the calculations could be accelerated from the outset of the conformational search procedure. The bond-scaling relaxation method[17,18] used by Zheng et al. randomly generated a number of conformations and subsequently scaled them to meet the distance constraints. The random configuration is minimized in the presence of bulk protein environment with the equilibrium bond lengths gradually restored during the minimization. The procedure is efficient in calculation. Monte Carlo "importance sampling" has been widely used in the computational procedure for determining the minimum energy conformation of protein molecules, which has been widely assumed to require an exponential amount of time with respect to the protein size, i.e., a NP-hard problem.[25] Carlacci and Englander[19] used a Monte Carlo algorithm to generate conformations for local segments in bovine pancreatic trypsine inhibitor (BPTI). In their approach, the computed loop segment started from a random conformation and was allowed to move while the rest of the protein remained fixed. Dummy residues identical in type and conformation with the residues on the fixed part of the protein immediately adjacent to the first and last residues of the segment were added to the ends of the segment. The total energy of the sampled conformations is the conformational energy of the loop segment in a local region plus a polypeptide chain continuity constraint represented by a harmonic overlapping energy of the dummy residues with the matching residues in the protein. The best conformation was chosen on the basis of the lowest total energy and was refined further. The Monte Carlo simulations of Higo and co-workers[20-22] started from an extended conformation of the loop with one terminal connected to the fixed bulk protein, then closed the loop by applying an harmonic potential to the four backbone atoms of the last residue of the loop to match the relative positions in experimental structure. Another important energy optimizing protocol, molecular dynamics, was also used in searching the conformational space of the protein loop[26,27] or cyclic peptide,[29] whose ability to reconstruct protein loops, however, still needs to be explored.

Monte Carlo method has been developed to be a mature and general algorithm in combinatorial optimization problems. There is a large variety of versions present for us to select. Another advantage of the Monte Carlo method is that it can efficiently search the conformational space and can almost always find low energy conformations (although probably not the lowest energy conformation), while the bond-scaling relaxation method sometimes cannot overcome local energy barriers during the minimization and fail to relax.[18] The Monte Carlo method can also easily be extended from single loop modeling to multiple loops modeling, as illustrated in the work of Higo and co-workers.[20,21] What they did was to add an extra term in the energy function representing interactions between multiple loops, then near-native conformations could be generated and selected by energy criteria. With the bond-scaling relaxation method, Rosenbach and Rosenfeld[29] used a simultaneous closure procedure to model two loops in a protein together. They first calculated a small number of putative starting conformations for the first loop, followed by simultaneous closure of each of these with 100 random starting conformations for the second loop. It was demonstrated that conformations close to the crystal structure could be generated for both loops, but there was no proper criteria to choose them, so it remains a problem as to whether this method could be extended to the case of more than two loops.

The main drawback of the Monte Carlo method in modeling protein loops and other macromolecule systems is its low efficiency. Most work adopted various improved versions of Monte Carlo algorithm to do the simulation of protein loop, such as the simulated annealing method used by Carlacci and Englander,[19] which has drawn much attention in the field of global optimization and was used, for example, to solve the traveling salesman problem[30]—one of the best-known NP-complete problems. Simulated annealing relies on a Monte Carlo procedure but, instead of being carried out at constant temperature, the simulation starts from a high temperature at which the system can overcome energy barriers and explore the configuration space widely. The temperature is slowly decreased, then it becomes increasingly difficult to cross energy barriers. Finally, it is hoped that the system is trapped in its ground state. This method takes its name from analogy with the metallurgical process of annealing. If a crystal solid is melted and then cooled too quickly (quenched), usually a disordered structure trapped in a local minimum results. If the system is annealed, i.e., brought to higher temperature and then slowly cooled, the crystalline ground state can be obtained. Higo and co-workers[20-22] developed an extended simulated annealing process by combining simulated annealing with the scaled collective variables method of Noguti and Go,[31] which can model the protein loops with a higher efficiency. Given the knowledge of the energy or statistical properties of conformational subspaces (e.g., $\phi$-$\psi$ zones or side-chain torsion angles), the biased probability Monte Carlo (BPMC) procedure designed by Borchert and coworkers[23,32] randomly selects the subspace, then takes a step to a new random position independent of the previous position, but according to the predefined continuous probability distribution. The random step is followed by a local minimization in torsion angle space. The BPMC runs displayed a much better convergence properties than the nonbiased simulations. Another method used by Vasmatzis et al.[24] can generate small, eight-backbone atom, local moves in Cartesian coordinates within geometric constraints, and then are efficient to be used in adaptable Monte Carlo procedures.

Although different algorithms were tested, the limit of computer power remains the main prohibitive factor for the Monte Carlo method to model long loops or multiple loops, because large conformational space need to be explored. In some cases, for example, when modeling CDRs in antibodies, it is better to consider six loops simultaneously if a more precise and reasonable result is expected. Most of the present Monte Carlo methods became time-consuming even when treating middle-sized loops (around 7 residues) on supercomputers[19] and parallel computers,[20,22] let alone longer ones. Therefore, algorithms with higher speed and efficiency will be helpful and necessary for treating long loops and multiple loops. As is well known, generally the limiting factor in a Monte Carlo calculation is the evaluation of potential energy. In loop modeling, people have to calculate both the intraaction within the loop segment and the interaction between the loop segment and the bulk protein environment in computationally acceptable time. Therefore, the key point in our approach is to introduce a potential in Monte Carlo algorithm, which should be effective enough in generating "correct" conformations on the one hand, and be simple in form suitable for fast computing on the other. In the following sections, we will describe how a simplified soft-sphere potential combined with a grid-mapping method has been successfully used to satisfy the dual purposes.

After generating a number of candidate conformations, the next problem is how to select near-native conformations from other putative conformations. The empirical energy is the criteria most often adopted, but generally it was considered dissatisfying. In this work, we also want to test if the empirical energy can work successfully in our method.

## METHODS

### Energy Function

In our MCSA simulation, the total conformational energy consists of two parts: the nonbonded atomic interaction represented by the soft-sphere potential and the harmonic distance constraint to ensure the smooth connection of the loop segment to the rest of the protein structure.

***Soft-Sphere Potential.*** Both experimental and theoretic work show that proper steric packing is the necessity of correct protein fold. Early calculations employing only the hard-sphere potential in which there is no attractive term and only an infinite repulsion when two atoms approach each other within a distance less than or equal to the sum of their van der Waals radii have been able to provide very useful, although approximate, insights into the structure of the oligopeptide chain.[33] Here we designed a so-called soft-sphere potential, as shown in Eq. (1), to calculate the nonbonded interaction between the nonbonded atomic pairs, which only evaluates the steric interaction by considering the van der Waals volume of atoms. We have successfully used the same potential in modeling the carbohydrate part of glycoproteins under the full bulk protein environment.[34]

$$E_s = \begin{cases} k_s(d_0^2 - d^2), & d_0 > d \\ 0, & d_0 < d \end{cases} \tag{1}$$

where $E_s$, $d$, $d_0$, and $k_s$ represent, respectively, the soft-sphere interaction energy between two atoms, the distance between them, the standard van der Waals distance between them that is equal to the sum of the standard van der Waals radii of atoms,[35] and the force constant that could be properly set by users. There is no special amendment considered to compensate the lack of hydrogen atoms in the simulation. The interactions for sequential atoms, i.e., 1-2 and 1-3 interactions, were omit-

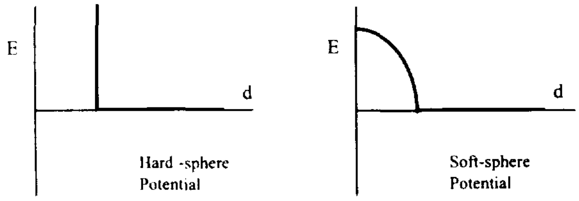

**FIGURE 1** Schematic drawing, comparing the hard-sphere and soft-sphere potentials. The potential energy $U$ is plotted as a function of the interatomic distance.

ted, and there is no difference in treating the interactions between 1-4 and 1-5 or the longer ones. This type of potential will neglect the interaction between atoms when their van der Waals volumes do not overlap. Clash is allowed, but the closer the atomic pair is, the more repulsive the interaction. It is displayed in Figure 1 that in contrast to the hard-sphere model with absolute exclusion, the soft-sphere model is "soft" and therefore can easily be used by the conformational searching methods, such as Monte Carlo and molecular dynamics, to scan a large conformational space.

This potential omitted the attractive term and static electric term to avoid artificial attractive forces between the atoms that could, in the absence of solvent molecules, lead to a biasing of conformations toward those that have internal van der Waals type attractions, i.e., conformations that show artificially high internal interactions.[4] The simple soft-sphere potential could be computed much faster than the complex 6-12 Lennard-Jones potential function, which is most often used in empirical force fields to represent the van der Waals interaction between nonbonded atomic pairs.

***Terminal Constraint.*** Apart from the nonbonded soft-sphere interaction, a harmonic distance constraint was used to ensure the smooth connection of the flexible loop segment to the rest of the protein structure fixed in their known structure during the simulation. The importance of anchorage in determining a strained protein loop conformation has been demonstrated by the experimental work of Hodel et al.[36] In the modeling work of Higo and his co-workers,[20-22] they first moved the N-terminal residue of the peptide segment to their reference position in the x-ray structure by appropriate translations and rotations, then constrained the main-chain atoms of the last residue of the segment to the same atoms in the x-ray structure. This measure is thought to be not very strict, as one cannot be sure that the terminal residues of any loop segment will be in the same position as the crystal structure before starting the modeling. Actually, their simulation of, for example, a 7-residue loop, only corresponds to the simulation of a 5-residue loop in other articles. Thus, the prediction accuracy in their papers may be overestimated when compared with other works. Carlacci and Englander[19] added two dummy residues to the two ends of the loop segment identical in type and conformation with the residues in the fixed part of the protein immediately adjacent to the first and last residues of the segment, then constrained them to their real positions in x-ray structure. This method was first proposed by Chou et al.[37] Although it is more reasonable than what Higo et al. did, we note that a neglect remains. Assuming a dummy residue with known conformation to the crystal structure was added to the C-terminal of the loop segment illustrated in Figure 2, one has to assign a definite $\phi$ angle for the residue; otherwise, there will be infinite number of possible connected conformations for the dummy residue, as shown in Figure 2. But if one sets the angle according to its value in the crystal structure, one takes for granted that the position of the carbonyl carbon atom of the last residue of the loop segment is already known, which is the only way the $\phi$ angle could be determined. This is not strict because we should ensure that we have no prior knowledge of the detailed positions of loop atoms at the beginning of an ab initio procedure. Alternatively, one can add a rotation freedom to the algorithm letting the $\phi$ angle in Figure 2 variable in the simulation, but an extra computational effort will be brought in.

From Figure 2, we can see that the key to ensure the smooth connection of two residues is to keep the atoms N and C$\alpha$ of one residue in the same *trans* peptide plane with the atoms C$\alpha$, C, and O of its preceding residue, i.e., the $\omega$ angle should be 180° or 0°. From this geometric relation, the coordinates of the N and C$\alpha$ atoms of a residue can be calculated from its preceding one given the parameters of the standard bond length and bond angle.[38] Based on this, we presented a more strict procedure to anchor the loop segment to its neighbor residues. First, two dummy atoms N and C$\alpha$ were generated following the fixed part residue that is right adjacent to the N-terminal of the loop segment. Next, a stretched peptide was generated by CHARMm as the loop segment. Then, the peptide was translated and rotated to have the two atoms N and C$\alpha$ of its N-terminal residue overlap with the dummy atoms, i.e., to be anchored to the fixed part of protein. Finally, another two dummy atoms N and C$\alpha$ were generated to connect to the last residue of the loop peptide to work as a constrained target in the Monte Carlo simulation. The constrained energy was represented by a harmonic overlapping function of the latter two dummy atoms from their reference positions in the crystal structure:

$$E_c = k_c\{[\mathbf{r}(\mathrm{N}) - \mathbf{r}_0(\mathrm{N})]^2 + [\mathbf{r}(\mathrm{CA}) - \mathbf{r}_0(\mathrm{CA})]^2\} \quad (2)$$

where $\mathbf{r}(\mathrm{X})$ and $\mathbf{r}_0(\mathrm{X})$ are the position vectors of a dummy atom X and its reference atom in the crystal structure; $k_c$ is the force coefficient.

***Total Energy.*** The total conformational energy was the sum of the nonbonded soft-sphere energy and the constrained energy for loop closure. No bonding energy terms (bond-stretch, bond angle bending, torsion angle rotating, etc.) were considered.

$$E = E_s + E_c \tag{3}$$

The soft-sphere energy was calculated for all nonbonded atomic pairs both within the flexible loop segment and between the loop segment and the bulk protein. The two environmental atoms N and C$\alpha$ that are immediately next to the C-terminal of the loop segment were not involved in the computing of the nonbonded interaction with the loop segment. Because the bulk protein part was fixed during the simulation, the nonbonded interactions within them should be constant and therefore were not calculated.

## Methods to Accelerate the Calculation of Environmental Interactions

The number of nonbonded atomic pairs within a loop segment is generally quite smaller than the number of those between loop segment and its environment, so the key factor responsible for the speed of the algorithm will be the calculation of the environmental interaction. We take a 5-residue loop in BPTI, as an example, which typically consists of 40 atoms. After randomly rotating a torsion angle, to evaluate the energy variation we need to recalculate less than 40 nonbonded interactions within the loop segment for each rotated loop atom. Comparatively, the calculation of the environmental interaction are rather time-consuming, as there are 412 atoms out of the totally 454 atoms in BPTI taken as the environmental atoms for the 5-residue loop (412 = 454-40-2, two atoms N and C$\alpha$ immediately next to the C-terminal of the loop were not involved in the computing), i.e., 412 nonbonded environmental interactions need to be recalculated for each rotated atom if no proper accelerating method is introduced.

***Local Region Method.*** Distance cutoff was often used in the empirical force field calculations to lighten the computing burdens, as in the work of Higo et al.[20-22] Carlacci et al. used an ellipsoid to define a local region; the residues of the bulk protein enclosed in the ellipsoid were involved in the computing of nonbonded interaction.[19] At the beginning of our work, we similarly tried the local region method by designing a special ellipsoid. Since the clear definition of the ellipsoid was not given in Carlacci et al.'s work,[19] here we strictly devise the ellipsoid in Figure 3. First, a referential ellipsoid was constructed, whose two focuses are just the two anchor atoms immediately adjacent to the two terminals of loop segment, i.e., the carbonyl carbon atom C adjacent to the N-terminal and the nitrogen atom N- to the C-terminal. The sum of the distances of any point on the ellipsoid surface to the two focuses should be a constant, and here it is set to the main-chain length of the stretched loop segment. Therefore, the ellipsoidal surface is approximately the outermost position that the main-chain atoms of the loop segment can reach. To correctly estimate the number of environmental atoms possible that clash with the main-chain atoms, the ellipsoidal volume should expand outside a specified distance $D_c$, which is the distance that two atoms start to clash, i.e., the sum of the van der Waals radii of two atoms. Likewise, one more expansion of a proper value $D_s$ for the space of the loop side chain was necessary. Then finally, a new ellipsoid was used to approximate the local region of the environment, whose intercepts in the three axes of the Cartesian coordinate system each has an increment of $D_c + D_s$ to the referential ellipsoid. The advantage of using an ellipsoid to define the local region is the ease of judging whether a protein atom falls into the local region: if the distance sum of this atom to the two focuses of the ellipsoid is longer than the specified distance constant for the surface point of the ellipsoid, the atom will be undoubtedly outside of the local region. The focuses of the new ellipsoid are $F_C$ and $F'_N$ shown in Figure 3. In the simulation, we only need calculate the interaction inside of the ellipsoid local region to evaluate the influence of the environment.

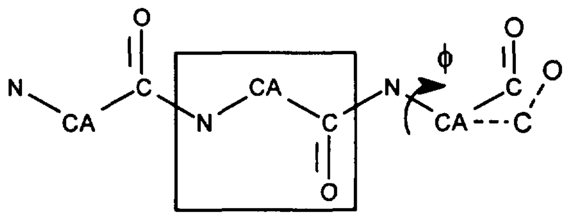

FIGURE 2 The method of adding dummy residue and dummy atoms. Inside of the square box is the loop segment (as an example, only one residue was demonstrated); the dashed line conformation represents an alternative conformation possibly adopted by the dummy residue with a different $\phi$ angle.

***Grid-Mapping Method.*** In order to further increase the computing speed, we developed a grid-mapping method to accelerate the calculation of the environmental interactions. First, we generated a cuboid that could exactly include the van der Waals volume of the whole fixed protein part. The cuboid was divided into a large number of small cubes with a grid size $D_g$, an example of which was shown in the center of Figure 3 drawn in heavy lines. Randomly putting an atom inside of the cube, we can see that any other atoms that possibly clash with this atom should locate at least within the outside polyhedron. The figure only shows the upright part of the polyhedron facing the reader for an easy view; the rest can be deduced based on symmetry. In our program, we recorded the bulk protein atoms that possibly clash with every cube in a special array, i.e., the bulk protein atoms were mapped onto every cube. To evaluate the interaction of a flexible loop atom with the bulk protein environment, we only need to determine, first of all, which

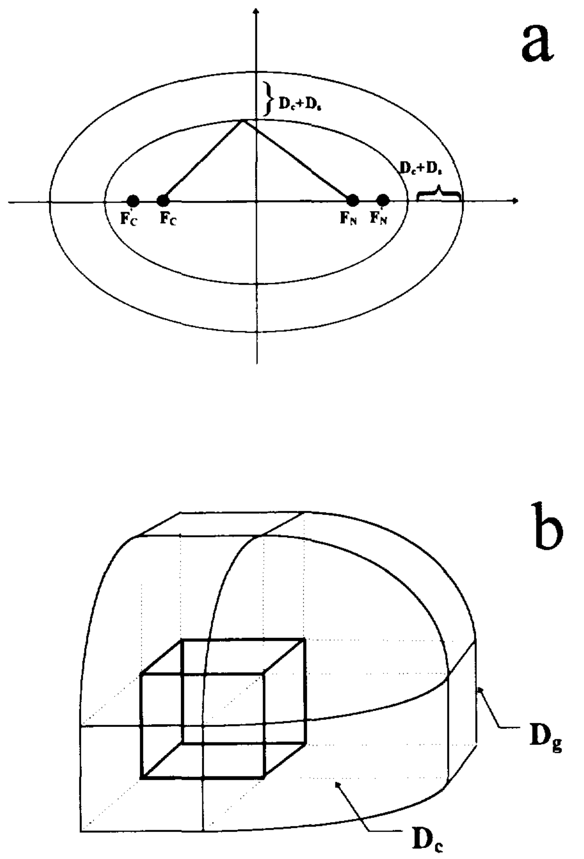

FIGURE 3 Methods to accelerate the calculation of the environmental interaction. (a) Local region method and (b) grid-mapping method. $D_c$, $D_s$, and $D_g$ are double of van der Waals radii of sulfur atom, side-chain length, and grid size. $F_C$, $F_N$ and $F'_C$, $F'_N$ are the focuses of the referential ellipsoid and the local region ellipsoid, respectively.

cube it locates in, then calculate the energy of this atom to all the environmental atoms mapped onto this cube.

The grid-mapping method can determine all the possible nonbonded atoms around a loop atom definitely, without the defect of the local region method of ignoring some of the possible environmental atoms. The grid size can be scaled to decrease the polyhedron volume, and consequently to reduce the number of environment atoms wherein. The volume of the polyhedron in Figure 3 can be represented as

$$V_p = D_g^3 + 6D_g^2 D_c + 3\pi D_c^2 D_g + \tfrac{4}{3}\pi D_c^3 \qquad (4)$$

The grid size, on the other hand, cannot be too small. Otherwise, too many grids will cause the computer memory overflow. Thus, a trade-off needs to be made based on the local hardware situation. In our simulation, it was set to 1 Å; then the volume is 340 $Å^3$ for BPTI, which at most can hold about 24 atoms for the atomic density of protein BPTI, estimated to be 14–18 $Å^3$ per atom by means of dividing the total atom number of BPTI (454) into its whole volume. Consequently, the number of environmental atoms is about the same as the number of atoms in a loop segment, and even less in case of long loops. The real number of environmental atoms is actually smaller than the estimated value of 24, because the cubes in which the loop atoms are possibly located are around or outside of the protein surface and hence do not have many environmental atoms concomitant. Therefore, the huge amount of computing effort for evaluating the environmental interaction can be reduced dramatically.

## MCSA Algorithm

In a MCSA run, all the protein atoms except the loop atoms were kept fixed in the same coordinates as in the crystal structure. The method to generate initial conformation of the loop segment has been described in the previous section. All the variable torsion angles, including main chain $\phi,\psi$ angles and side-chain $\chi$ angles, were set to 180° with the exception of the $\phi$ angle of proline, which was fixed at 75° during the simulation. Conformations were sampled by randomly selecting one of the variable torsion angles and assigning a new value between −180° and +180°. Main-chain $\omega$ angles were kept to be 180° during the simulation. The total energy of the new conformation, which consists of the nonbonded soft-sphere energy and the terminal constrained energy, was evaluated in each step and compared to the prior one; then the new conformation was either accepted or rejected based on the Metropolis criteria.[39] A cycle was completed after a specified number of conformations are searched. The number was set to 100 times the number of variable torsion angles. The accepting rate of each cycle was equal to the number of accepted conformation divided by the total conformational search number. The simulation starts from a high temperature where a high accepting rate (80% in our work) can be achieved, and the temperature was lowered at the end of each cycle by multiplying a scale factor of 0.83. When the temperature dropped to a value near zero, or when the simulation run a specified number of cycles, the simulation was terminated. The uniform random number generator GGL[40] based on a linear congruential method was adopted in our MCSA. The whole algorithm was implemented in C++ code.

## Simulation Procedures

The protein structure of BPTI (PDB code 4PTI) from the PDB was chosen as an example in the simulation, which is also the most frequently used protein structure in testing loop modeling methods. The same segments as calculated by Carlacci and Englander[19] were used here, i.e., loops of LP1(12–16), LP2(11–17) and LP3(10–18) and $\alpha$-helix of LP4(46–50), $\beta$-strand of LP5(16–

20), the numbers in parentheses indicating residue sequential numbers. In order to remove bad contacts and to compare the calculated result with the crystal structure, first the crystal structure of 4PTI was optimized with 200 steps of steepest descents (SD), 200 steps of conjugate gradient (CONJ), and 800 steps of Adopted Basis Newton-Raphson (ABNR) method successively using CHARMm (QUANTA 4.0)[41] with polar hydrogens. Since it is hard to give a precise dielectric constant for the surface part of proteins, we set it to the value in vacuum, i.e., one unit, for simplicity. The energy coefficients in the soft-sphere potential and the constrained potential are 10 and 100 kcal/(mole atom Å$^2$), respectively. The computer we used was SGI Indy/R4400 with

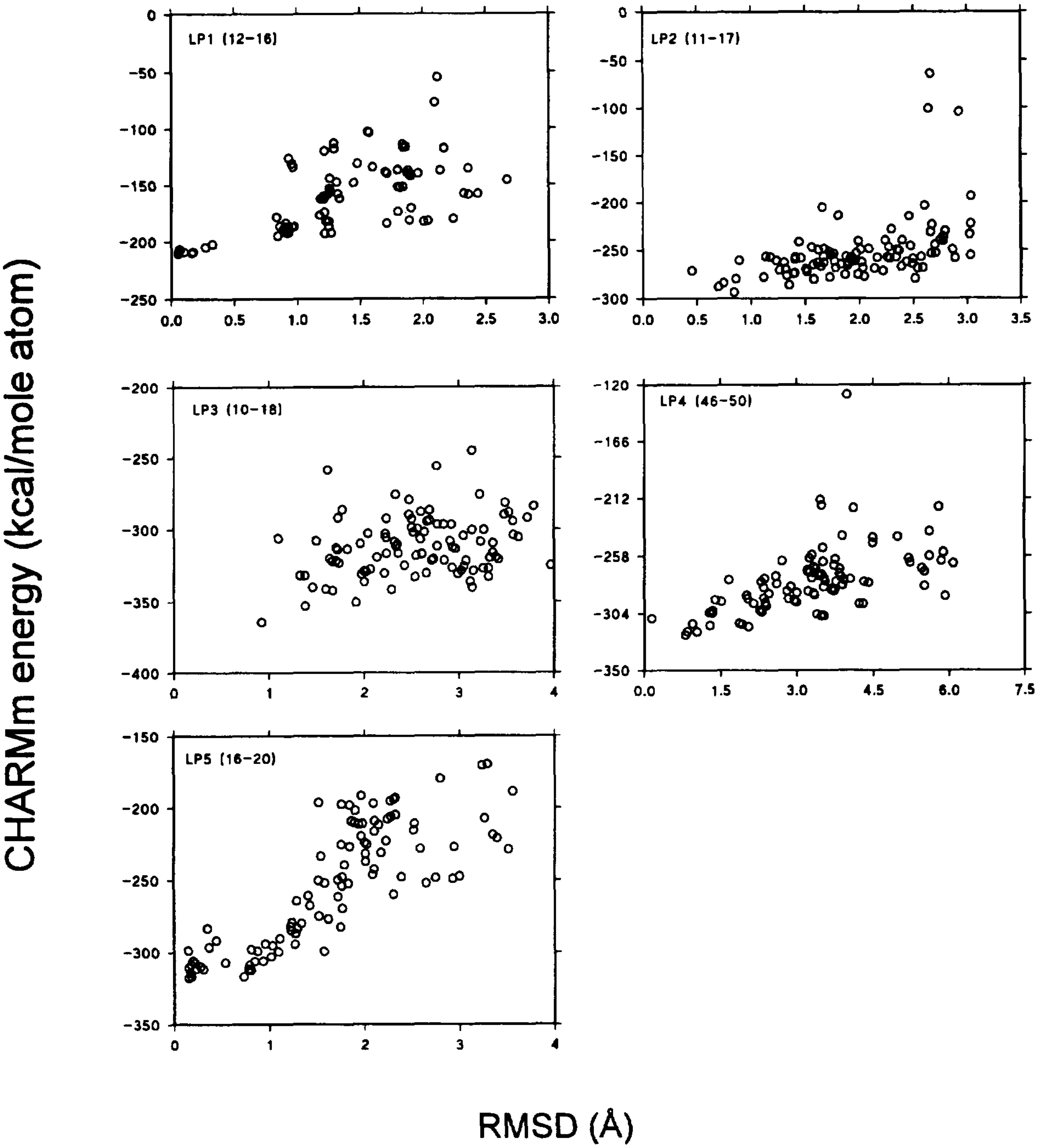

FIGURE 4 CHARMm energy vs RMSD from crystal structure for the conformations generated by the local region method. There is one data point out of the figure range in LP1 and LP3 respectively because of its high energy and therefore is not shown here.

**Table I RMSDs from Crystal Structure of the Predicted Conformations for the Five Loops in BPTI (Å)**

| | Backbone[a] | | | | All atoms | | | |
|---|---|---|---|---|---|---|---|---|
| | Lowest Energy Conformation | | Lowest RMSD Conformation | | Lowest Energy Conformation | | Lowest RMSD Conformation | |
| Loop Segment | L[b] | G[c] | L[b] | G[c] | L[b] | G[c] | L[b] | G[c] |
| LP1 (12–16) | 0.15 | 0.06 | 0.06 | 0.04 | 0.53 | 0.34 | 0.25 | 0.34 |
| LP2 (11–17) | 0.84 | 0.84 | 0.45 | 0.70 | 1.90 | 1.90 | 1.78 | 1.78 |
| LP3 (10–18) | 0.93 | 0.93 | 0.93 | 0.93 | 2.34 | 2.34 | 2.34 | 2.34 |
| LP4 (46–50) | 0.80 | 0.88 | 0.14 | 0.06 | 2.62 | 1.96 | 1.45 | 1.54 |
| LP5 (16–20) | 0.73 | 0.26 | 0.15 | 0.12 | 1.23 | 1.95 | 1.15 | 1.02 |

[a] Including main chain atom N, $C\alpha$, C, and O.
[b] Result of the local region method.
[c] Result of the grid-mapping method.

174 MHZ IP22 processor, 48 megabytes main memory and operating system of IRIX5.3.

With each accelerating method (local region and grid mapping), 100 candidate conformations were generated by the MCSA program for every loop. Those conformations were then optimized with 400 steps of the SD, CONJ, and ABNR method successively using CHARMm. During the minimization, all protein atoms except the loop atoms are kept fixed in their crystal position; therefore, the CHARMm energy consists of the energy inside of the loop segment and the energy between the loop segment and its environment.

## RESULTS

What we are mostly concerned are, first of all, whether we can get near-native conformations and select them with a proper criteria; second, whether we can obtain the results with high speed and efficiency.

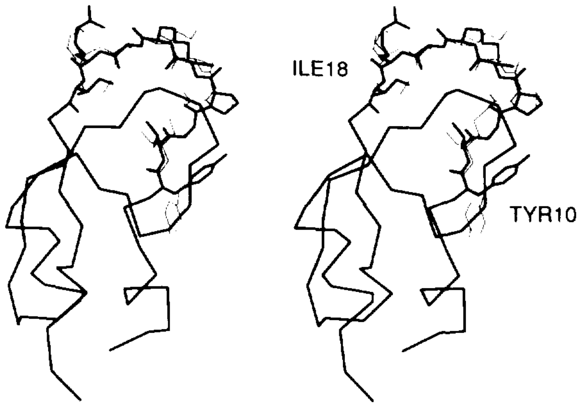

FIGURE 5 The structure of the longest 9-residue loop LP3. Light line is the simulated conformation with RMSD of 0.93 Å and the heavy line is the crystal structure.

### Modeling Precision

We calculated the rms deviations (RMSDs) of the 100 generated conformations from the crystal structure for every loop and correlated them with their CHARMm energies. The result of the local region method is displayed in Figure 4. It can be seen that there is a large number of conformations reaching a high precision to the crystal structures, for example, the RMSDs from the crystal structure of many of them are even below 0.1 Å for 5-residue loops. It was shown in Table I that all the lowest energy and lowest RMSD conformations have RMSDs from the crystal structure below 1 Å, and those generated by the grid-mapping method are slightly better than those by the local region method in most cases.

Although it was thought that the empirical energy was not a very good criteria in judging the correctness of loop folding and some improved free energy functions have been suggested,[42] we still use it in our work with a moderate success (shown in Figure 4 and Table I); worthy of mention is the correct selection of the lowest backbone RMSD conformation of 0.93 Å for the longest 9-residue loop LP3. The simulated conformation and the crystal structure of LP3 are displayed in Figure 5.

### Speed of the Program

In the local region method, the volume of the ellipsoid depends on two variables: $D_t$ and $D_s$. In the

**Table II The Number of Atoms in Local Regions (when $D_c$ and $D_s$ in Figure 3 are 3.6 and 3 Å, respectively)**

| Segment | LP1 | LP2 | LP3 | LP4 | LP5 |
|---|---|---|---|---|---|
| Loop | 31 | 49 | 69 | 37 | 43 |
| Environment | 208 | 343 | 377 | 336 | 191 |

simulation, $D_c$ was set to be double the van der Waals radius of sulfur atom, i.e., 3.6 Å, and the side-chain expanding $D_s$ was set to a tentative value 3 Å, which is only proper for short and middle-length side chains. The numbers of the environmental atoms within the local region for the five-loop segments range from 191 to 377 (shown in Table II), which are similar to those in Ref. 19. The data suggest that the method can only save a limited amount of computing time, i.e., about half of a plain method or less, because totally we have around 400 bulk protein atoms. In case of the longest 9-residue loop, almost all of them were included; thus very limited computational time could be saved with the local region method. We can reduce the number of the environmental atoms by decreasing the value of $D_s$ to contract the size of the ellipsoid; but actually, even the original value of 3 Å is not big enough for most residues with long or even middle side chains. Therefore, the local region method worked not so successfully as we hoped.

The grid-mapping method, however, worked very successfully, as demonstrated by the dramatic reduction of the program running time shown in Table III. Actually, the grid-mapping method cost less than one-tenth time as much as the local region method.

The step of optimization with CHARMm costs around 1-2 min for each conformation, which can generally decrease 1-2 Å of the RMSD value.

## DISCUSSION

The MCSA method based on the soft-sphere potential was demonstrated to be quite efficient in loop modeling, which suggests that the steric constraint and the geometric constraint are the most important factors for the correct orientation of protein loops. More explicit potential could be used at the final step as the tool of refinement.

Comparing the results here to the previous reference, we found that all the lowest energy RMSDs are better than those in Ref. 19, and some of them are nearly 1 Å lower than those in the reference. But this comparison is not stringent enough because of the different regulation strategy of the initial crystal structure in the two studies.

The local region method based on the commonly used distance cutoff technique was not very successful in accelerating the computation of environmental interactions, while the grid-mapping method could dramatically reduce the time. In the grid-mapping method, although the grid size could be reduced even smaller than its present value of 1 Å if a computer with a larger memory were available, no obvious increasing of the computational speed might be expected, because the time for calculating the interactions within the loop segment already is comparable to the time for calculating the environmental interactions when the grid size was 1 Å, i.e., the key to influence the computational time is no longer only the interactions from the environment.

Starting from the same set of random numbers, in most cases with two accelerating methods we got the same resultant conformations, especially in long loops, as seen in Table IV. There are two extreme cases in the table: LP3 and LP5. In the former, almost all resultant conformations are the same in two methods (96%), while none is the same in the latter. This can be explained by the size

**Table III Typical Running Time of a Single MCSA Procedure**

| | | Running Time (min) | |
|---|---|---|---|
| Loop Segment | Residue/Torsion Angle Number | Local Region Method | Grid-Mapping Method |
| LP1 | 5/14 | 18 | 1-2 |
| LP2 | 7/24 | 93 | 5 |
| LP3 | 9/32 | 150 | 12 |

**Table IV Comparison of the RMSDs from Crystal Structure of the 100 Conformations Generated by the Two Accelerating Methods Based on the Same Series of Random Seeds**

| RMSD Comparison | LP1 | LP2 | LP3 | LP4 | LP5 |
|---|---|---|---|---|---|
| =[a] | 44 | 84 | 96 | 70 | 0 |
| >[b] | 30 | 9 | 1 | 17 | 51 |
| <[c] | 26 | 7 | 3 | 13 | 49 |

[a] The number of conformations having the same RMSDs in two methods.

[b] The number of conformations having the larger RMSDs in the local region method than in the grid-mapping method.

[c] The number of conformations having the smaller RMSDs in the local region method than in the grid-mapping method.

of the local regions. It was shown in Table II that the local region of LP3 is so large that almost all the bulk protein atoms were included. Therefore, the resultant conformations calculated by the local region method are mostly the same as those by the grid-mapping method. In case of LP5, the number of the environmental atoms in the local region is quite low, making the energy hypersurface in the local region method obviously different from the real one in the grid-mapping method. In most loops, the conformations generated by the local region method have larger RMSDs more often than those by the grid-mapping method, as shown in Table IV. An exception is LP3, for which we generated two more near-native conformations by the local region method than by the grid-mapping method. But so few data points (4) cannot assure a statistical conclusion. Then, generally, the grid-mapping method can have a higher accuracy than the local region method, because it determines the environment of loop atoms, definitely avoiding the necessary approximation adopted in the local region method.

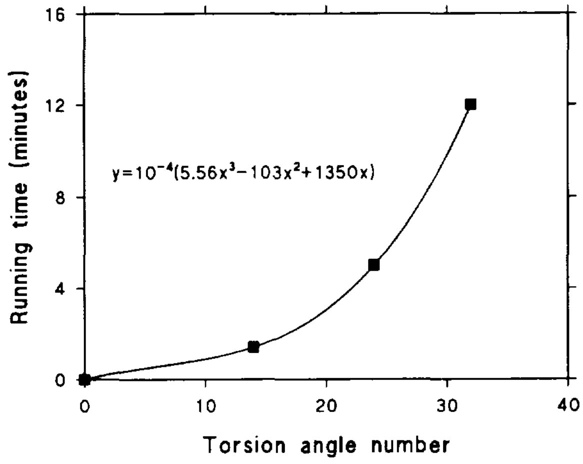

FIGURE 6 The correlation between the running time and the number of the variable torsion angles of the MCSA program with grid-mapping method.

Similar to the previous work, the empirical energy function cannot always successfully select the lowest RMSD conformation. But in our work, the lowest energy conformations are very close to the lowest RMSD conformations in all cases, which proved that, as a quite useful tool, the empirical energy function still can be effectively used in judging the correctness of the fold of local segments.

Finally, we can see from Figure 6 that the relation between the computational time of the MCSA program with the grid-mapping method and the number of torsion angles can be approximately fitted by a cubic curve, while in the systematic algorithm the relation is the function of $e^n$, where $n$ is the system size. Therefore, the advantage of the Monte Carlo method in solving the combinatorial problems makes possible the use of modeling of even longer loops or multiple loops in the near future. Furthermore, more fast and efficient computations could be achieved if more advanced versions of MCSA algorithms in Refs. 20 and 23 were adopted.

The authors thank the Chinese State Commission of Science and Technology, State Education Commission, and the Nature Science Foundation for financial support. We also thank Professor Kankaala of Tampere University of Technology of Finland for providing the random number generator code.

*Note.* We have uploaded the source code of our program in the anonymous FTP server of Institute of Physical Chemistry (IPC) of Peking University to make our method easily to be reproduced by others. The server address is: ftp.ipc.pku.edu.cn., with path: /pub/upload/lpsa/.

## REFERENCES

1. Rees, A. R., Staunton, D., Webster, D. M., Stearte, S. J., Henry, A. H. & Pedersen, J. T. (1994) *TIBTECH* **12,** 199–206.

2. Vasquez, M., Nemethy, G. & Scheraga, H. A. (1994) *Chem. Rev.* **94,** 2221–2239.
3. Padlan, E. A. & Kabat, E. A. (1991) *Methods Enzymol.* **203,** 3–21.
4. Martin, A. C. R., Cheetham, J. C. & Rees, A. R. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 9268–9272.
5. Jones, T. A. & Thirup, S. (1986) *EMBO J.* **5,** 819–822.
6. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987) *Nature* **326,** 347–352.
7. Moult, J. & James, M. N. G. (1986) *Proteins Struct. Func. Genet.* **1,** 146–163.
8. Sudarsanam, S., Dubose, R. F., March, C. J. & Srinivasan, S. (1995) *Protein Sci.* **4,** 1412–1420.
9. Chothia, C., Lesk, A. M., et al. (1989) *Nature* **342,** 877–883.
10. Reczko, M., Martin, A. C. R., Bohr, H. & Suhai, S. (1995) *Protein Eng.* **8,** 389–395.
11. Go, N. & Scheraga, H. A. (1970) *Macromolecules* **3,** 178–187.
12. Bruccoleri, R. E. & Karplus, M. (1985) *Macromolecules* **18,** 2767–2773.
13. Bruccoleri, R. E. & Karplus, M. (1987) *Biopolymers* **26,** 137–168.
14. Dudek, M. J. & Scheraga, H. A. (1990) *J. Comput. Chem.* **11,** 121–151.
15. Palmer, K. A. & Scheraga, H. A. (1991) *J. Comput. Chem.* **12,** 505–526.
16. Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. & Levinthal, C. (1987) *Biopolymers* **26,** 2053.
17. Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1993) *J. Comput. Chem.* **14,** 556–565.
18. Zheng, Q., Rosenfeld, R., Vajda, S. & DeLisi, C. (1993) *Protein Sci.* **2,** 1242–1248.
19. Carlacci, L. & Englander, S. W. (1993) *Biopolymers* **33,** 1271–1286.
20. Higo, J., Collura, V. & Garnier, J. (1992) *Biopolymers* **32,** 33–43.
21. Gibrat, J., Higo, J., Collura, V. & Garnier, J. (1992) *Immunomethods* **1,** 107–125.
22. Collura, V., Higo, J. & Garnier, J. (1993) *Protein Sci.* **2,** 1502–1510.
23. Borchert, T. V., Abagyan, R., Radha Kishan, K. V., Zeelen, J. P. & Wierenga, R. K. (1993) *Structure* **1,** 205–213.
24. Vasmatzis, G., Brower, R. & Delisi, C. (1994) *Biopolymers* **34,** 1669–1680.
25. Ngo, J. T. & Marks, J. (1992) *Protein Eng.* **5,** 313–321.
26. Karplus, M. & Bruccoleri, R. E. (1990) *Biopolymers* **29,** 1847–1862.
27. Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L. & Levinthal, C. (1986) *Proteins Struct. Funct. Genet.* **1,** 342–362.
28. Mao, B., Maggiora, G. M. & Chou, K. C. (1991) *Biopolymers* **31,** 1077–1086.
29. Rosenbach, D. & Rosenfeld, R. (1995) *Protein. Sci.* **4,** 496–505.
30. Rees, S. & Ball, R. C. (1987) *J. Phys. A Math. Gen.* **20,** 1239–1249.
31. Noguti, T. & Go, N. (1985) *Biopolymers* **24,** 527–546.
32. Borchert, T. V., Abagyan, R., Jaenicke, R. & Wierenga, R. K. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1515–1518.
33. Scheraga, L. (1992) in Lipkowitz, K. B. & Boyd, D. B., Eds., *Reviews in Computational Chemistry,* Vol. 3, VCH, New York, pp. 73–142.
34. Zhang, H., Yang, Y., Lai, L. & Tang, Y. (1996) *Carbohydr. Res.,* **284,** 25–34.
35. Bondi, A. (1964) *J. Phys. Chem.* **68,** 441–451.
36. Hodel, A., Kautz, R. A., Adelman, D. M. & Fox, R. O. (1994) *Protein Sci.* **3,** 549–556.
37. Chou, K. C., Nemethy, G., Pottle, M. S. & Scheraga, H. A. (1985) *Biochemistry* **24,** 7948–7953.
38. Zhang, H. Doctoral Thesis, Peking University, 1996.
39. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. J. (1953) *J. Chem. Phys.* **21,** 1087–1092.
40. Lewis, P., Goodman, A. & Miller, J. (1969) *IBM Sys. J.* **2,** 136–159.
41. QUANTA4.0 manual, Molecular Simulation Incorporated, 1994.
42. Smith, K. C. & Honig, B. (1994) *Proteins Struct. Funct. Genet.* **18,** 119–132.
43. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.