

# The Loop Problem in Proteins: A Monte Carlo Simulated Annealing Approach

LOUIS CARLACCI\* and S. WALTER ENGLANDER

The Johnson Research Foundation, Department of Biochemistry and Biophysics, University of Pennsylvania, School of Medicine, Philadelphia, Pennsylvania 19104-6059

## SYNOPSIS

A Monte Carlo simulated annealing (MCSA) algorithm was used to generate the conformations of local regions in bovine pancreatic trypsin inhibitor (BPTI) starting from random initial conformations. In the approach explored, only the conformation of the segment is computed; the rest of the protein is fixed in the known native conformation. Rather than follow a single simulation exhaustively, computer time is better used by performing multiple independent MCSA simulations in which different starting temperatures are employed and the number of conformations sampled is varied. The best computed conformation is chosen on the basis of lowest total energy and refined further. The total energy used in the annealing is the sum of the intrasegment energy, the interaction energy of the segment with the local surrounding region, and a distance constraint to generate a smooth connection of the initially randomized segment with the rest of the protein.

The rms deviations between the main-chain conformations of the computed segments in BPTI and those of the native x-ray structure are 0.94 Å for a 5-residue  $\alpha$ -helical segment, 1.11 Å for a 5-residue  $\beta$ -strand segment, and 1.03, 1.61, and 1.87 Å for 5-, 7-, and 9-residue loop segments. Side-chain deviations are comparable to the main-chain deviations for those side chains that interact strongly with the fixed part of the protein. A detailed view of the deviations at an atom-resolved level is obtained by comparing the predicted segments with their known conformations in the crystal structure of BPTI. These results emphasize the value of predetermined fixed structure against which the computed segment can nest.

© 1993 John Wiley & Sons, Inc.

## INTRODUCTION

We are concerned with the structure of loop regions defined in the broadest sense as the conformation of any continuous chain segment within a known protein structure. The problem of defining an unknown loop conformation within a known protein structure occurs in many guises, for example, the hypervariable loops in antibodies,<sup>1-3</sup> the loops connecting the seven-helix-bundle of membrane-bound bacteriorhodopsin,<sup>4</sup> the changes that may result from mutations within a local region of a known protein structure,<sup>5,6</sup> and the transiently formed

nonhydrogen-bonded loops that appear to underlie protein hydrogen exchange processes.<sup>7,8</sup>

These and other cases would profit from a computational methodology that could help define the unknown conformation of a loop when the protein structure it nests against is known. In the simulations tested here, the conformation of a segment of a protein is computed while the rest of the protein remains fixed in the initial conformation. The power of the method depends on knowledge of the part of the protein structure that does not change.

As a first step we explore the ability of Monte Carlo simulated annealing (MCSA) to reproduce the known native conformation of various segments ( $\alpha$ -helical,  $\beta$ -strand, loop) in bovine pancreatic trypsin inhibitor (BPTI). In each calculation, the conformation of the segment is initially randomized and then refolded in a conformational search for the lowest energy structure. The MCSA procedure used

\* Present address is Department of Chemistry, University of South Florida, Tampa, FL 33620-5250.

to refold the segment allows the system to pass through high as well as low energy pathways. As part of the annealing process, the temperature is lowered stepwise so that the refolding is sequentially restricted to lower energy pathways. Rather than follow a single simulation exhaustively and impose some convergence test, computer time is utilized to perform multiple independent simulations of limited extent. The "best" simulation is then selected on the basis of lowest computed energy among the several attempts and refined further. For the present simulations of known structures, the degree of success can be judged by comparison with the known native structure. The rms deviations of main-chain heavy atoms obtained between the computed and x-ray conformations are 0.9–1.1 Å for various types of 5-residue segments and 1.6–1.9 Å for 7- and 9-residue loops. Side chains that interact strongly with the fixed part of the protein achieve about the same degree of success while freely exposed side chains do not.

In previous work, a number of groups have attacked the problem of predicting the conformations of local regions in proteins. Brucoleri and Karplus devised a method for uniformly sampling the conformational space of short polypeptide segments in which various filters are used in the sampling.<sup>9</sup> Levinthal and collaborators<sup>2,10</sup> have implemented methods that cycle between energy minimization and molecular dynamics to predict the conformations of surface loops beginning from many initial conformations. For many arbitrary initial loop conformations in model protein tertiary folds, Chou et al.,<sup>11</sup> Carlacci and Chou,<sup>12,13</sup> and Chou and Carlacci<sup>14</sup> employed energy minimization to optimize the packing. In another study, Carlacci et al.<sup>15</sup> used a heuristic approach employing energy minimization to predict the conformations of large loops (40, 24, and 9 residues) in bovine somatotropin.

Simulated annealing (SA) approaches have been often used before to overcome the local minimum problem.<sup>16–22</sup> Lee and Subbiah<sup>16</sup> and Lee and Levitt<sup>17</sup> have recently used a MCSA method to predict side-chain conformations for proteins whose main-chain conformation is fixed in the known native conformation. In studies on open polypeptide chains, Wilson and Cui<sup>18</sup> found that MCSA alone without any restraints revealed the lowest energy conformation while using less computer time than rigorous conformational search algorithms. The SA method has been widely employed in calculations to refine x-ray<sup>19,20</sup> or nmr<sup>21,22</sup> models of proteins. However, the SA approach has not yet been used to predict the detailed conformations of initially random loops.

It appears that the common denominator in prior successful applications is the use of more or less known docking constraints. The calculations described here carry this a step further by initially randomizing both the main chain and side chains of a segment in a protein and then allowing the segment to search for its correct (known) conformation, docked against the rest of the protein, which remains fixed during the MCSA conformational search procedure.

## METHODS

This section describes the calculation to duplicate segments in bpti. The calculation defines the segment to be duplicated and residues that directly interact with it, randomizes the conformation of the segment, and then uses a MCSA refolding algorithm to search for its lowest energy conformation.

### Local Regions

We define the local region to consist of the residues around the segment to be computed and the segment itself. Energy calculations involve only the residues in the local region. The parts of the protein outside of the local region do not enter the calculation.

Local regions are generated by centering an ellipsoidal volume over the native conformation of the polypeptide segment to be computed. The shape of the ellipsoid is such that the segment and its nearby residues are contained inside. A residue is defined to be part of the local region when at least one of its atoms lies inside the volume. In these calculations the local region included residues within 7–8 Å from the computed segment (see Table I).

The segments in BPTI that are replicated in the calculations are a 5-residue  $\alpha$ -helical segment located at the N-terminus of the helix, a 5-residue  $\beta$ -strand segment, and 5-, 7-, and 9-residue loop segments. The segments' names and a description of their local regions are given in Table I. As an example, the name B5 (16–20) corresponds to a 5-residue  $\beta$ -strand segment in BPTI that includes residues 16–20. The computed loop segments are centered on Cys14, which forms a disulfide bridge with Cys38. The disulfide bridge constraint<sup>23</sup> restricts the movement of the chain, and thereby reduces the computational effort required to predict the loops. The disulfide bridge plays a smaller role for the 7- and 9-residue loop segments and is not present in the helical and  $\beta$ -strand segments. A stereo stick model of BPTI that highlights the  $\alpha$ -helical segment,  $\beta$ -strand segment, and the 9-residue loop segment

**Table I** The BPTI Segments Studied<sup>a</sup>

	H5(46-50)	B5(16-20)	L5(12-16)	L7(11-17)	L9(10-18)
Type	$\alpha$ -Helix	$\beta$ -Strand	Loop	Loop	Loop
Local region <sup>b</sup>	4, 19-23, 30-33, 42-45, 51-55	9-12, 14-15, 21-22, 32-38, 40, 44-46	10-11, 17-20, 33-41, 44	8-10, 18-22, 31-47	8-9, 19-22, 32-46

<sup>a</sup> In each named segment, the numbers of the first and last residues are indicated in parentheses. The amino acid sequences of the computed segments are as follows:

B5(16-20): Ala-Arg-Ile-Ile-Arg  
H5(46-50): Lys-Ser-Ala-Glu-Asp  
L5(12-16): Gly-Pro-Cys-Lys-Ala  
L7(11-17): Thr-Gly-Pro-Cys-Lys-Ala-Arg  
L9(10-18): Tyr-Thr-Gly-Pro-Cys-Lys-Ala-Arg-Ile

Disulfide bonds exist between residues 5-55, 14-38, and 30-51. Data were taken from entry 4pti<sup>34</sup> in the Protein Data Bank<sup>35,36</sup> at Brookhaven National Laboratory.

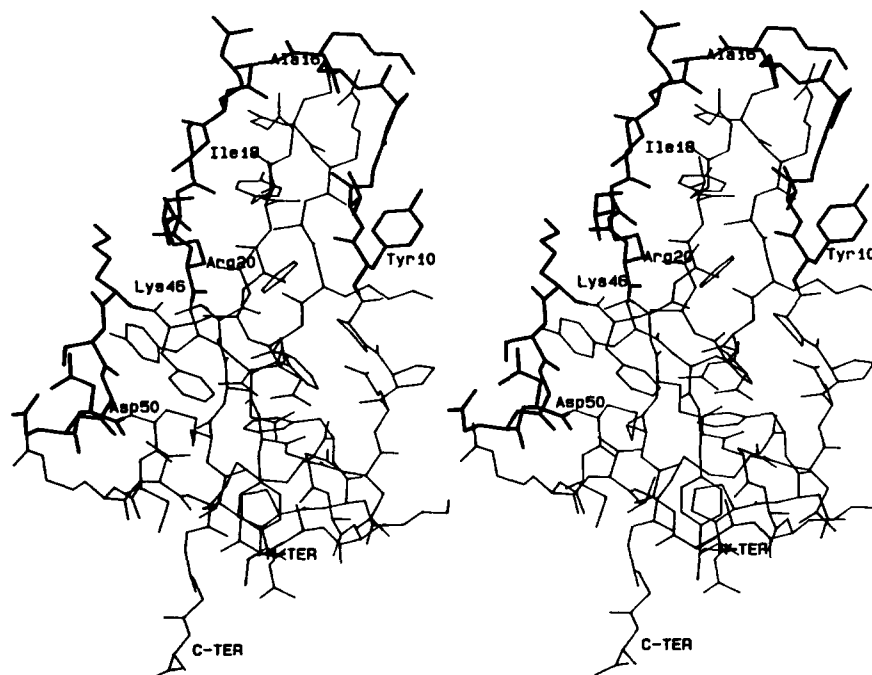
<sup>b</sup> The residues in the local region that surround the computed segment. Energetics between the segment and residues in the local region that surround it are included in the total energy that is evaluated in the annealing.

is depicted in Figure 1. Amino acid sequences of the segments are listed in the footnotes to Table I.

### Coordinates of the Computed Segment

During a simulation, the computed segment is allowed to move while the rest of the protein remains

fixed. Dummy residues identical in type and conformation with residues on the fixed part of the protein immediately adjacent to the first and last residues of the segment are added to the ends of the segment. When the segment to be computed is initially randomized, the segment is severed from the protein. The purpose of the dummy residues is to



**Figure 1.** Stick model representation of the native x-ray crystal structure of BPTI. Thick lines represent the parts of the protein that were refolded in this study. The labels indicate the first and last residues of segments computed [B5(16-20): Ala16, Arg20; H5(46-50): Lys46, Asp50; L9(10-18): Tyr10, Ile18]. Only heavy atoms are represented.

force the detached segment to smoothly reconnect with the rest of the protein by overlapping them with the matching residues in the protein during the simulation.

Two types of coordinates are needed to define the conformation of a segment in a polypeptide chain, namely local dihedral angles and rigid body coordinates. All bond lengths and bond angles are fixed at standard values.<sup>23</sup> For each segment computed, there are 6 rigid body coordinates, including 3 Euler angles— $\alpha$ ,  $\beta$ , and  $\gamma$ —which orient the segment, and 3 translational coordinates— $T_x$ ,  $T_y$ , and  $T_z$ —which position the segment in the polypeptide chain. The Cartesian coordinates of the segment are computed from dihedral angles by considering the segment an independent polypeptide chain in its own reference frame, called  $\mathbf{r}'$ . In the  $\mathbf{r}'$  reference frame, the peptide N of the first residue of the segment, i.e., the N-terminus dummy residue in all the segments computed here, defines the origin. The initial orientation of the segment is such that the C $^\alpha$  of the first residue

is positioned out along the positive  $x$  axis and the terminal NH is positioned in the  $xy$  plane along the positive  $y$  direction.

The coordinates of the segment and the dummy residues are transformed to the coordinate system of the protein by use of Eq. (1), which orients and positions the segment and dummy residues in the polypeptide chain.

$$\mathbf{r}(W) = \Omega \mathbf{r}'(W) + \mathbf{T} \quad (1)$$

In Eq. (1),  $\mathbf{r}'(W)$  and  $\mathbf{r}(W)$  denote two types of position vectors for atom  $W$  in the segment. The  $\mathbf{r}'(W)$  is the position of atom  $W$  in the segment computed from dihedral angles. The  $\mathbf{r}$  without a prime denotes the reference frame where the segment is in the protein;  $\mathbf{r}(W)$  is the position vector of atom  $W$  in the protein. Equation (1) is described in Ref. 24 for the case of separate polypeptide chains. In Eq. (1),  $\mathbf{T}$  is a translation vector and  $\Omega$  is a unitary matrix computed from Euler angles as in Eq. (2):

$$\Omega = \begin{bmatrix} \cos \alpha \cos \gamma - \sin \alpha \cos \beta \sin \gamma & -\cos \alpha \sin \gamma - \sin \alpha \cos \beta \cos \gamma & \sin \alpha \sin \beta \\ \sin \alpha \cos \gamma + \cos \alpha \cos \beta \sin \gamma & -\sin \alpha \sin \gamma + \cos \alpha \cos \beta \cos \gamma & -\cos \alpha \sin \beta \\ \sin \beta \sin \gamma & \sin \beta \cos \gamma & \cos \beta \end{bmatrix} \quad (2)$$

The initial translation vector is just the position vector  $\mathbf{r}(N)$  of atom  $N$  in the first residue of the segment positioned in the polypeptide chain. The elements of the first column of  $\Omega$  are the coefficients of the unit vector along the bond connecting N to C $^\alpha$  in the  $\mathbf{r}$  reference frame. The elements of the second column are the coefficients of the unit vector obtained by a Schmidt orthogonalization procedure where the unit vector along the bond connecting N to H in the  $\mathbf{r}$  reference frame is made orthogonal to the first vector and normalized. The elements in the third column are the coefficients of the unit vector obtained from the cross product of the first two vectors.

### Empirical Energy Parameters

The empirical energy of a trial conformation is computed with ECEPP/2 (Empirical Conformational Energy of Polypeptides Program) geometric and energy parameters<sup>25</sup> updated from Momany et al.<sup>23</sup> In the ECEPP protocol, bond angles and bond lengths are fixed geometric parameters obtained from crystal structures of dipeptides. Also, the conformational energy  $E_{\text{conf}}$ , of a polypeptide chain is given by

$$E_{\text{conf}} = E_{\text{es}} + E_{\text{nb}} + E_{\text{hb}} + E_{\text{tor}} \quad (3)$$

where the terms are, respectively, the electrostatic, nonbonded, hydrogen-bonded, and torsional energy. The electrostatic energy is a typical Coulomb potential with a constant dielectric of 2. The nonbonded energy, the sum of atom pair interaction energies, is computed from the Lennard-Jones 6–12 potential function. Two types of nonbonded parameters are employed to compute the nonbonded energy, one for interactions between two atoms separated by exactly three bonds (1–4 interactions) and the other for interactions between two atoms separated by more than three bonds (1–5 interactions). The interaction energies between two atoms separated by one or two bonds are not computed since the bond lengths and angles associated with these interactions are fixed.  $E_{\text{hb}}$  has parameters for 1–4 and 1–5 interactions also, and uses the 10–12 form of the potential function. The torsional energy is given by a trigonometric function that depends on the number of minima that occur for rotation about the bond and the size of the energy barriers. The neutral forms of all ionizable groups on the amino acids are utilized in the calculation, and no explicit solvent is present.

To reduce the computational effort, united atom potentials are used in the energy calculations. For

example, in a CH<sub>3</sub> to CH<sub>2</sub> interaction, 12 pairwise interactions are replaced by a single interaction energy. Pairwise united atom potentials are derived from the all-atom set. Thus the united atom CH<sub>3</sub> to CH<sub>2</sub> pairwise potential is the sum of the 12 separate pairwise potentials. Also the united atom internuclear distance where the potential function is at a minimum is taken to be the same as for the all-atom C to C interaction. The net partial charge is positioned on C. For polar atoms and their hydrogens, the energy parameters are unchanged from the all atom set. All-atom potentials were used for energy minimization of the native x-ray structure and in the final simulation to refine the computed loop conformation having the lowest energy.

### The Total Conformational Energy of the Local Region

The total conformational energy ( $E_{\text{total}}$ ) evaluated to decide which of the sampled conformations to save or reject during the annealing simulation is the sum of the conformational energy of the local region ( $E_{\text{local}}$ ) plus a polypeptide chain continuity constraint. The total conformational energy is given by

$$E_{\text{total}} = E_{\text{local}} + \text{XLOOP} \times F_{\text{overlap}} \quad (4)$$

The conformational energy of the local region is given by

$$E_{\text{local}} = E_{\text{intra}}(\text{cs}) + E_{\text{inter}}(\text{cs/local}) \quad (5)$$

where  $E_{\text{intra}}(\text{cs})$  is the intrasegment energy of the computed segment, and  $E_{\text{inter}}(\text{cs/local})$  is the sum of the interaction energies between the computed segment and segments in the local region that surround it.<sup>12,26</sup> When a residue in the computed segment is adjacent to (shares a common peptide bond with) a residue of a segment in the local region, a distinction must be made between sequential and nonsequential segments. For nonsequential segments, the atom pair interactions are 1–5 interactions. For a segment in the local region that immediately precedes or follows a segment duplicated (sequential segments), only the pairwise interactions between the two residues that share the intersegment peptide bond need special consideration, namely, some pairwise interactions are of the 1–2, 1–3, and 1–4 type.<sup>12,26</sup> All the other pairwise interactions between the sequential segments are 1–5 interactions. The 1–2 interaction corresponding to the C'—N bond stretching and the 1–3 interactions corresponding to the bending of the bond angles  $\angle\text{C}^\alpha\text{—C'—N}$ ,  $\angle\text{O—C'—N}$ ,  $\angle\text{C'—N—C}^\alpha$ , and

$\angle\text{C'—N—H}$  are not computed. In the method developed here, these effects are built into a polypeptide chain continuity constraint. The intersegment energy contribution for sequential segments is then the sum of 1–4 and 1–5 interaction energies. Residues outside the local region make no contribution to  $E_{\text{local}}$ .

The other component of the total conformational energy of a local region,  $\text{XLOOP} \times F_{\text{overlap}}$ , is a polypeptide chain continuity constraint in the form of an harmonic potential that forces the randomized segment, covalently disconnected from the protein in the initial randomization process, to smoothly connect with the fixed part of the protein during the MCSA trajectory. For this purpose, the residues on the fixed part of the protein immediately adjacent to the first and last residues of the computed segment are called overlapping residues, and dummy residues identical in type and conformation with the overlapping residues are appended to the ends of the computed segment.<sup>11</sup> A smooth connection is reached when the overlapping residues and the dummy residues merge, i.e., when the bond length and angles associated with the initially broken peptide bond settle near their standard geometry values.<sup>23</sup> An alternative method would be to force the loop segment to be sequential with the rest of the protein for each conformation generated by the Monte Carlo algorithm, as in the chain closure algorithm of Go and Scheraga.<sup>27</sup> By allowing the end points of the segment to detach from the rest of the protein and by allowing all specified degrees of freedom to vary independently, the method we employ gives the segment much more freedom to sample conformations, especially at high temperatures.

The overlap function ( $F_{\text{overlap}}$ ) is the sum of the squares of the distances between the N, C<sup>α</sup>, C', and O backbone atoms in the dummy residues and the corresponding atoms in the overlapping residues [Eq. (6)].<sup>11</sup>

$$F_{\text{overlap}} = \sum [\mathbf{r}_i(\text{N}) - \mathbf{r}_i^0(\text{N})]^2 + [\mathbf{r}_i(\text{C}^\alpha) - \mathbf{r}_i^0(\text{C}^\alpha)]^2 + [\mathbf{r}_i(\text{C}') - \mathbf{r}_i^0(\text{C}')]^2 + [\mathbf{r}_i(\text{O}) - \mathbf{r}_i^0(\text{O})]^2 \quad (6)$$

In Eq. (6),  $\mathbf{r}_i(X)$  is a position vector for atom  $X$  in the  $i$ th dummy residue and  $\mathbf{r}_i^0(X)$  is a position vector for the identical atom in the overlapping residue. Based on trial and error, a force constant of 100 kcal/(mol Å<sup>2</sup>) appears to work well for XLOOP, and is used in the calculations reported here.

The value of the rms overlap deviation indicates how well the segment connects with the fixed part

of the protein. From  $F_{\text{overlap}}$  in Eq. (6), the rms overlap deviation is given by

$$\text{rms overlap} = [F_{\text{overlap}}/8]^{1/2} \quad (7)$$

where a total of 8 atoms in the overlapping residues are overlapped by the 2 dummy residues appended to the segment.

### Monte Carlo Simulated Annealing Algorithm

The formulation and implementation of MCSA has been extensively documented.<sup>18,28-32</sup> This discussion describes the protocol followed here. In brief, a typical simulated annealing run starts at high temperature, and the temperature is slowly lowered at intervals throughout the simulation. Conformations are sampled by randomly picking one of the coordinates that define the segment duplicated and assigning a new value to it. The total energy for the new conformation is evaluated and compared to the prior one. The new conformation is accepted or rejected based on the Metropolis criteria.<sup>28</sup> A step in the simulation is completed after a specified number of conformations are evaluated. The temperature is lowered at the end of each step. After a specified number of steps, the temperature is set to zero and the last step of the simulation is executed.

In the SA calculation to compute the conformation of the segment, the coordinates that are parameters are the Euler angles and translational coordinates, all the side-chain dihedral angles and the  $\phi$  and  $\psi$  main-chain dihedral angles<sup>33</sup> of all residues except for  $\phi$  of proline that, according to the ECEPP protocol,<sup>23</sup> is fixed at  $-75^\circ$ . In the final simulation to refine and optimize the computed segment the  $\omega$  dihedral angles, otherwise set at  $180^\circ$ , are allowed to vary.

To initially randomize the segment to be computed, values of the coordinates are randomly chosen according to

$$\begin{aligned} -180^\circ < \mathbf{w} &\leq 180^\circ \\ -180^\circ < \Gamma &\leq 180^\circ \\ \mathbf{T}^\circ - 1.0 \text{ \AA} &\leq \mathbf{T} \leq \mathbf{T}^\circ + 1.0 \text{ \AA} \end{aligned} \quad (8)$$

Here  $\mathbf{w}$  represents the dihedral angle parameters in the annealing, and  $\Gamma$  represents the Euler angles.  $\mathbf{T}$  is the translation vector and  $\mathbf{T}^\circ$  is the initial translation vector before randomization.

Conformational space is sampled based on a random walk. One of the coordinates is randomly chosen and assigned a random value according to

$$C_{\text{old}} - \text{pert} \leq C_{\text{new}} \leq C_{\text{old}} + \text{pert} \quad (9)$$

where  $C_{\text{old}}$  is the prior value,  $C_{\text{new}}$  is the new value, and  $\text{pert}$  is a perturbation defined for each type of coordinate (dihedral angle, Euler angle, translational) as summarized below.

Parameters that define the annealing schedule are the temperature ( $RT$ ), temperature reducing factor (TFAC), number of conformations sampled per step (NCON), and number of steps (NSTEP). Temperature is referred to in units of  $RT$ , where  $R$  is the gas constant in kcal/(mol Kelvin), and  $T$  is the temperature in Kelvin.  $RT$  is a measure of the barrier size that can be overcome during the simulation. The factor to reduce the temperature at the end of each step in the simulation is computed according to

$$\text{TFAC} = \{RT_{\text{init}}/RT_{\text{NSTEP}-1}\}^{\text{NSTEP}-2} \quad (10)$$

where  $RT_{\text{init}}$  is the initial temperature and  $RT_{\text{NSTEP}-1}$  is the desired temperature just before the last step.

The rate of cooling can be adjusted by  $RT_{\text{init}}$ , TFAC, and NSTEP. In our calculations, TFAC is computed and NSTEP is the same for all simulations (29 steps from  $RT_{\text{init}}$  to  $RT_{29} = 0.6$  kcal/mol, plus 1 step at  $RT_{30} = 0$  kcal/mol, for a total of 30 steps). Different rates of cooling are achieved by independently varying  $RT_{\text{init}}$ . Under these conditions, the larger the value of  $RT_{\text{init}}$ , the larger is the temperature reducing factor. Another consequence of different values of  $RT_{\text{init}}$  is that the random walk pathway for refolding the segment is different. This is especially true for the large variation in  $RT_{\text{init}}$  tested here. The initial temperatures employed are based on preliminary calculations.

A new conformation is accepted or rejected depending upon the Metropolis criteria.<sup>28</sup> A new conformation with a total energy less than or equal to that for the prior conformation is always accepted as a conformation from which to begin the next sampling cycle. When the new conformation has a larger total energy, the probability of accepting the higher energy conformation is the Boltzmann probability of making the jump in energy. According to the Metropolis criteria a new conformation is accepted when

$$\text{RAND} \leq p \quad (11)$$

where RAND is a number randomly chosen in the range of 0 to 1, and  $p$  is the probability of accepting the new conformation. The probability is given by

$$p \Leftarrow \begin{cases} 1 & (\text{when } \Delta E_{\text{total}} \leq 0) \\ \exp[-\Delta E_{\text{total}}/RT] & (\text{when } \Delta E_{\text{total}} > 0) \end{cases} \quad (12)$$

where  $\Delta E_{\text{total}}$  is the difference in total energy between the new conformation and the prior conformation, and  $RT$  is the temperature in the current step of annealing. If the ratio of the number of accepted conformations in a step to the number NCON (acceptance ratio) at the end of a step falls below 25%, the values of the perturbations used in Eq. (9) are reduced by a factor of two. The smaller steps enhance the probability for the system to overcome energy barriers.<sup>16,29</sup>

The parameters for the annealing are summarized as follows:

*Annealing Schedule to Compute the Segments*

$$\text{pert} = \begin{cases} 90^\circ \text{ for dihedral angles} \\ 45^\circ \text{ for Euler angles} \\ 1.0 \text{ \AA for translations} \end{cases}$$

$$RT_{\text{init}} = 5000, 2000, 1000, \text{ and } 500 \text{ kcal/mol}$$

$$RT_{\text{NSTEP}-1} = 0.6 \text{ kcal/mol}$$

(corresponding to 300 K)

$$\text{NSTEP} = 30 \text{ steps}$$

$$\text{acceptance ratio} = 25\%$$

*Annealing Schedule to Refine the Segments*

$$\text{pert} = \begin{cases} 10^\circ \text{ for dihedral angles} \\ 5^\circ \text{ for Euler angles} \\ 0.1 \text{ \AA for translations} \end{cases}$$

$$RT_{\text{init}} = 1.0 \text{ kcal/mol; restart at step 29 of best}$$

(lowest energy) conformation computed

$$RT_{\text{NSTEP}-1} = 0.2 \text{ kcal/mol}$$

$$\text{NSTEP} = 15 \text{ steps}$$

$$\text{acceptance ratio} = 25\%$$

### Selection of the Final Conformation

Multiple simulations are run for each segment computed, and the conformation with the lowest total energy is selected for further study. The parameters that differ for each simulation are the number of conformations (NCON) sampled per step and the initial temperature. For a given value of NCON, the several initial temperatures constitute a group of simulations (see Table II as an example). A group

of simulations is computed for each of two values of NCON, one much larger than the other. If the final total energy is significantly lowered with the larger value of NCON, NCON is increased again and another group of simulations is computed. The conformation with the lowest total energy at the end of the MCSA calculation is selected and refined further.

### Native and Refined Native X-Ray Structure of BPTI

The x-ray crystal structure named 4pti<sup>34</sup> in the Protein Data Bank<sup>35,36</sup> is taken as the native (x-ray crystal) structure. The refined native structure of BPTI is the result of two calculations; the first computes an optimized set of dihedral angles from the x-ray coordinates based on distance optimization and the second is an energy minimization. The distance optimization and energy minimization use a general unconstrained optimizing algorithm.<sup>37</sup>

In the first calculation, the dihedral angles are optimized to bring the computed coordinates as close as possible to the x-ray coordinates. Initial dihedral angles are computed from the heavy atom coordinates in the native x-ray structure; bond lengths and angles have standard values.<sup>23</sup> A distance optimization is employed in which the object function is the sum of the squares of the distances between the heavy atoms in the computed structure and the corresponding atoms in the x-ray structure.

The structure computed from the x-ray coordinates is then refined to remove bad contacts by four steps of energy minimization. Dihedral angles of side chains that could not be computed because one of the atoms is a hydrogen (for example,  $\chi_1$  of alanine) were assigned values from a look-up table.<sup>38</sup> These dihedral angles are optimized first. The parameters optimized in the next three energy minimizations are (1) all the side-chain dihedral angles, (2) all the  $\phi$ ,  $\psi$ , and side-chain dihedral angles, and (3) all dihedral angles including  $\omega$ . In the energy minimization, all-atom potentials are employed to compute the conformational energy of the polypeptide chain.

### Performance of the Program

The MCSA program, called SAFD, was installed and optimized on the CRAY Y-MP/832 at the Pittsburgh Supercomputer Center. Optimizing the program increased the MFLOPS rate by 6.1-fold from 13 to 79. However, the processor time required to do these calculations was only reduced by a factor of 2.7. The reason for the disagreement is that the MFLOPS rate, the sum of the number of floating

**Table II** Segment B5(16–20) in BPTI Duplicated by MCSA<sup>a</sup>

Run <sup>b</sup>	Annealing Schedule			RMS Deviations <sup>i</sup>					
	RT <sup>c</sup>	TFAC <sup>d</sup>	NCON <sup>e</sup>	$E_{\text{local}}^f$	$E_{\text{total}}^g$	Overlap <sup>h</sup>	Backbone <sup>j</sup>	Side chain <sup>k</sup>	All <sup>l</sup>
minU	—	—	—	−116.2	−116.2	0.0	0.14	0.18	0.16
initU	—	—	—	$3.7 \times 10^5$	$4.7 \times 10^5$	11.08	7.17	10.00	8.47
1	5000	1.38	12K	−29.9	164.8	0.493	1.72	1.92	1.81
2	2000	1.34	12K	103.4	211.2	0.367	2.16	7.41	5.07
3	1000	1.30	12K	100.6	197.3	0.348	1.57	5.19	3.57
4	500	1.27	12K	132.3	272.9	0.419	2.05	5.89	4.12
5	5000	1.38	18K	−11.4	44.8	0.265	1.66	3.67	2.69
6	2000	1.34	18K	−83.6	−37.9	0.239	2.69	7.34	5.17
7	1000	1.30	18K	−104.5	−96.5	0.100	1.16	3.64	2.51
8	500	1.27	18K	46.4	295.8	0.558	2.08	6.33	4.39
9	5000	1.38	24K	−97.2	−77.7	0.156	1.16	3.47	2.42
10	2000	1.34	24K	2.7	121.1	0.385	2.44	6.88	4.83
11	1000	1.30	24K	−16.9	123.5	0.419	1.99	6.17	4.27
12	500	1.27	24K	105.8	300.6	0.494	2.02	6.30	4.36
minA	—	—	—	−74.2	−74.2	0.0	0.14	0.18	0.16
7RA	1.0	1.13	18K	−63.6	−61.0	0.057	1.11	2.91	2.06

<sup>a</sup> See Table I for a description of the computed segments.<sup>b</sup> The number or name given to the simulation run. A run that begins with min means refined (by energy minimization) native x-ray conformation, a run that begins with init means initial randomized conformation, R is a MCSA refinement run, U and A indicate that the energetic analysis was made using united-atom or all-atom potentials, respectively. (See Methods for details of calculations and parameters employed.)<sup>c</sup> RT is the initial temperature in kcal/mol.<sup>d</sup> The temperature reducing factor used in each step. Number of steps (NSTEP) is 30.<sup>e</sup> The number of conformations computed per step, where 1K means 1000 conformations.<sup>f</sup> The energy of the local region in kcal/mol, i.e., the sum of intrasegment and intersegment energies.<sup>g</sup> The total energy, i.e., the sum of the energy of the local region and the weighted polypeptide chain continuity constraint.<sup>h</sup> The rms overlap deviation (in Ångströms) between backbone atoms (N, C $\alpha$ , C', and O) in the two dummy residues appended to the ends of the computed segment and the same atoms in the overlapping residues on the fixed part of the protein.<sup>i</sup> The rms deviations between the computed segment and the x-ray conformation (heavy atoms, in Ångströms). All heavy atoms in the protein except for those in the computed segment are superposed on the x-ray structure. (For all computed segments described in Table I, rms deviations between atoms superposed are backbone, < 0.18 Å; side chain, < 0.24 Å; all, < 0.20 Å).<sup>j</sup> Main-chain deviation between the computed and x-ray conformation (atoms N, C $\alpha$ , C $\beta$ , C', O, and C $\delta$  of proline).<sup>k</sup> Deviation between side-chain heavy atoms, except C $\delta$ , in the computed and x-ray conformation of the segment.<sup>l</sup> Deviation between all heavy atoms in the computed and x-ray conformation of the segment.

point multiplications, divisions, and additions divided by the number of seconds of processor time, does not include instructions such as logical statements. In a typical simulation, the program spends over 70% of the processor time in a single subroutine, namely the routine that computes the nonbonded, hydrogen-bonded, and electrostatic energies.

## RESULTS

### Calculation for a 5-Residue $\beta$ -Strand Segment

A typical example of a MCSA calculation is given by the calculation to compute the  $\beta$ -strand segment B5(16–20) (see Table I for a description). Aspects unique to longer loops will be addressed below. Results from multiple simulations to compute B5(16–

20) are given in Table II, where the first column is the label for the simulation run, columns 2–4 detail the annealing schedules employed, columns 5–7 characterize the energetics of the final conformation, and the last 3 columns give rms deviations of the computed segment from the native x-ray conformation.

The run labeled minU in Table II an energy minimization refinement calculated with all-atom potentials but evaluated here with united-atom potentials, gives the energetics for the local region of B5(16–20) in the refined native conformation that the computed segment should approach in the calculation. The entry labeled initU shows that the initial randomized conformation of the segment to be refolded is far from the native x-ray conformation.

Run initU is characterized by a high energy (computed using united-atom potentials), a large rms overlap deviation, 11.1 Å, since the segment no longer connects with the rest of the protein, and large rms deviations from the native x-ray conformation, e.g., 7.17 Å rms backbone atom deviation.

Results from three groups of refolding simulations are given in Table II. In these runs, the energy of each sampled conformation was evaluated using united-atom potentials. A given simulation was started at the indicated  $RT$  value (column 2), and a number of conformations were sampled (NCON, column 4). The temperature was then decreased by the factor, TFAC, shown in column 3, and another NCON conformations were generated at the new  $RT$ . TFAC is calculated to bring  $RT_{29}$  to 0.6 kcal/mol (room temperature) after 29 such steps. A final step involving NCON conformations was then performed at  $RT_{30} = 0$  kcal/mol. For the final conformation obtained at the end of each simulation, Table II lists the total energy ( $E_{\text{total}}$ ), the contribution to  $E_{\text{total}}$  due to the energy of the local region ( $E_{\text{local}}$ ), the rms deviation of the dummy residues from the overlapping residues on the fixed part of the protein, and the heavy atom rms deviations between the computed segment and the native x-ray segment. On the basis of energy, the best result for the series with NCON equal to 18K (run 7) is similar to the best 24K result (run 9). The final conformations from these lowest energy simulations also have the smallest rms deviations of the backbone from the native x-ray conformation.

A refinement simulation employing all-atom potentials to compute the energy was performed that started from the conformation at the end of step 29 of the simulation with the lowest total energy (simulation 7 in Table II). In the refinement simulation, a jump in the temperature to 1 kcal/mol was introduced, then the temperature was stepped down to 0.2 kcal/mol in 14 steps (NCON = 18K conformations at each value of  $RT$  as in run 7). A final step was run with  $RT_{15} = 0$  kcal/mol. The resulting conformation is characterized in the row marked 7RA (refined with all atom potentials). The refinement improves the final structure, apparently by removing some bad side-chain contacts (compare rms deviations for runs 7 and 7RA). The final conformational energy level reached compares well with the refined, energy-minimized native structure evaluated using all-atom potentials (minA). (Note that a meaningful energetic comparison between two conformations is possible when the energetics are computed with the same parameter set, i.e., united-atom or all-atom potentials.)

Deviations found between the computed B5 (16–20) conformation (7RA in Table II) and the native x-ray conformation are detailed in Table III. The total overall rms deviation of main chain atoms (N, C $^{\alpha}$ , C $^{\beta}$ , C $^{\gamma}$ , O, and C $^{\delta}$  of proline) from the native conformation is 1.11 Å. The total overall rms side chain deviation of 2.91 Å is largely contributed (53%) by the 6 atoms of the Arg17 side chain, which protrudes from the protein surface. When Arg17 is excluded, the rms deviation is 1.36 Å. The crystallographic temperature factors for the Arg17 guanidinium atoms<sup>34</sup> are very large compared to the rest of the atoms in the residue (C $^{\gamma} = 9.7$ , C $^{\delta} = 20$ , N $^{\epsilon} = 35$ , C $^{\zeta} = 31$ , N $^{\eta 1} = 31$ , and N $^{\eta 2} = 36$ ). In this study, predicted conformations of side chains that have x-ray temperature factors greater than about 25 usually show large computed deviations from the native x-ray conformations. Side chains in the computed segment that nest against the fixed part of the protein have much smaller overall rms deviations, in the range of 1–1.6 Å.

Figure 2 compares the computed B5 (16–20) segment with the native conformation. The deviation of the Arg17 side chain from the native conformation is clearly visible. The 1.8 Å backbone atom deviation of Ile18 reported in Table III points to another flaw; the hydrogen bond between the Ile18 carbonyl oxygen and the peptide hydrogen in the adjacent strand is broken in the computed conformation (Figure 2).

### Calculations for Other Segments

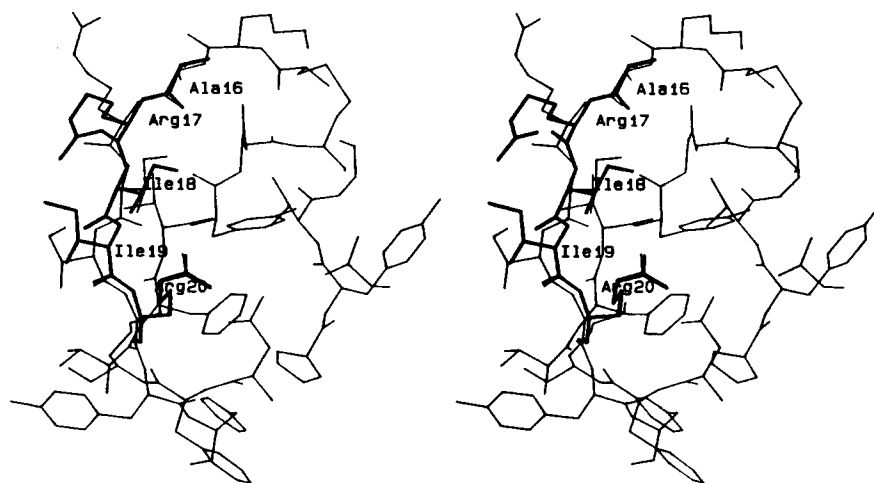
Other segments in BPTI, a 5-residue  $\alpha$ -helical segment and loop segments of 5, 7, and 9 residues, were computed as just described. The number of conformations sampled per step were 24, 36, and 42K for

**Table III** Deviations of B5(16–20) from Native X-Ray Conformation in Ångstroms<sup>a</sup>

	Backbone <sup>b</sup>	Side Chain <sup>c</sup>	All <sup>d</sup>
Ala16	0.51 (5)	— (0)	0.51 (5)
Arg17	1.26 (5)	4.66 (6)	3.55 (11)
Ile18	1.84 (5)	1.01 (3)	1.58 (8)
Ile19	0.90 (5)	1.61 (3)	1.22 (8)
Arg20	0.26 (5)	1.36 (6)	1.02 (11)
Total	1.11 (25)	2.91 (18)	2.06 (43)

<sup>a</sup> The refined lowest total energy conformation obtained from MCSA is analyzed here. Numbers in parentheses are the number of heavy atoms used to calculate the rms deviation. The rms deviations between atoms in the refined native segment and atoms in the x-ray conformation are backbone, 0.14 Å; side chain, 0.18 Å; all, 0.16 Å. See footnote i in Table II.

<sup>b–d</sup> See footnotes j–l in Table II.



**Figure 2.** Local region for the computed B5(16–20) segment. The computed segment is depicted by the thick line drawing, the native x-ray conformation of the local region is depicted by the thin line drawing. Only heavy atoms are represented.

segment H5(46–50), 12, 18, and 24K for segment L5(12–16), 18, 30, and 42K for segment L7(11–17), and 36, 42, and 54K for L9(10–18). The components of the total energies for the refined computed segments are given in Table IV, and the rms deviations from the native conformations are given in Tables V–VIII. These results are for the computed conformations with the lowest total energy after further MCSA refinement. The predicted conformations are compared to the native conformations in Figures 3 and 4a–c.

Table V compares the computed H5(46–50) segment with the native x-ray conformation. The rms deviation for 25 backbone atoms is 0.94 Å. Figure 3 shows the native x-ray conformation of the local region for the H5(46–50) segment and the segment computed. The large rms deviation in the main-

chain conformation of Ala48 involves breakage of the hydrogen bond to the Met52 peptide hydrogen. This discrepancy can be visualized in Figure 3. The residues with x-ray temperature factors greater than 25 are Lys46 and Asp50. A relatively large rms side-chain deviation occurs for Glu49 even though the x-ray temperature factors are about 10 for this side chain on the surface of the protein. In the x-ray crystal conformation, the Glu49 side chain appears to be hydrogen bonded to its own peptide hydrogen exposed at the N-terminus of the helix.

Results from simulations for the L5(12–16), L7(11–17), and L9(10–18) segments are given in Tables VI, VII, and VIII, respectively. These segments are centered on residue Cys14, whose side chain tethers the loop to the fixed part of the protein via a disulfide bridge with the side chain of residue

**Table IV** Energetics of Computed Segments and Refined Native Conformation

	H5(46–50)	B5(16–20)	L5(12–16)	L7(11–17)	L9(10–18)
Native <sup>a</sup>					
$E_{\text{total}}^b$	–59.76	–74.25	–6.06	–39.04	–55.97
Predicted <sup>c</sup>					
$E_{\text{local}}^d$	–51.44	–63.58	–8.28	–17.31	23.16
$E_{\text{total}}^b$	–50.43	–60.99	–4.49	–2.41	32.96
Overlap <sup>e</sup>	0.036	0.057	0.069	0.137	0.111

<sup>a</sup> Data for native x-ray conformation of segment after energy minimization refinement. The rms overlap is zero for the native conformation, i.e.  $E_{\text{total}} = E_{\text{local}}$ .

<sup>b</sup> See footnote g of Table II.

<sup>c</sup> Data for refined lowest total energy conformation from MCSA.

<sup>d</sup> See footnote f of Table II.

<sup>e</sup> See footnote h of Table II.

**Table V** Deviations of H5(46–50) from Native X-Ray Conformation in Ångströms<sup>a</sup>

	Backbone <sup>b</sup>	Side Chain <sup>c</sup>	All <sup>d</sup>
Lys46	0.89 (5)	1.46 (4)	1.18 (9)
Ser47	0.66 (5)	0.78 (1)	0.68 (6)
Ala48	1.75 (5)	— (0)	1.75 (5)
Glu49	0.26 (5)	3.78 (4)	2.53 (9)
Asp50	0.19 (5)	1.50 (3)	0.93 (8)
Total	0.94 (25)	2.47 (12)	1.60 (37)

<sup>a</sup> The rms deviations between atoms in the refined native segment and atoms in the x-ray conformation are backbone, 0.13 Å; side chain, 0.16 Å; all, 0.14 Å. See also footnote a of Table III.

<sup>b-d</sup> See footnotes j–l of Table II.

Cys38. As the size of the computed loop increases from 5 to 7 to 9 residues, the rms backbone deviation increases from 1.03 to 1.61 to 1.87 Å. Stereoscopic stick models in Figures 4a–c compare the computed loop segments with the native x-ray conformations. The large x-ray temperature factors observed for the side-chain atoms of residues Lys15 and Arg17 (most are greater than 25 and sometimes undefined) coincide with large rms side chain deviations seen in the tables. For segment L7(11–17), the small rms deviation of the Lys15 side chain (1.64 Å) does not match the large crystallographic temperature factors observed.

The characteristics of the calculation to compute segment L9(10–18) are not the same as those given in Table II for B5(16–20). The decision to end the calculation was based on the fact that the total energy did not improve when the number of conformations sampled per step was increased to 54K. The lowest total energy conformation was computed with 42K conformations per step. The simulation that

**Table VI** Deviations of L5(12–16) from Native X-Ray Conformation in Ångströms<sup>a</sup>

	Backbone <sup>b</sup>	Side Chain <sup>c</sup>	All <sup>d</sup>
Gly12	0.78 (4)	— (0)	0.78 (4)
Pro13	1.35 (6)	2.52 (1)	1.57 (7)
Cys14	0.81 (5)	2.18 (1)	1.16 (6)
Lys15	1.37 (5)	3.92 (4)	2.80 (9)
Ala16	0.34 (5)	— (0)	0.34 (5)
Total	1.03 (25)	3.48 (6)	1.79 (31)

<sup>a</sup> The rms deviations between atoms in the refined native segment and atoms in the x-ray conformation are backbone, 0.22 Å; side chain, 0.21 Å; all, 0.21 Å. See also footnote a of Table III.

<sup>b-d</sup> See footnotes j–l of Table II.

**Table VII** Deviations of L7(11–17) from Native X-Ray Conformation in Ångströms<sup>a</sup>

	Backbone <sup>b</sup>	Side Chain <sup>c</sup>	All <sup>d</sup>
Thr11	1.82 (5)	1.03 (2)	1.63 (7)
Gly12	1.04 (5)	— (0)	1.04 (4)
Pro13	1.76 (5)	1.61 (1)	1.74 (7)
Cys14	2.17 (5)	0.34 (1)	1.99 (6)
Lys15	1.22 (5)	1.64 (4)	1.43 (9)
Ala16	1.94 (5)	— (0)	1.94 (5)
Arg17	0.52 (5)	3.72 (6)	2.77 (11)
Total	1.61 (35)	2.66 (14)	1.97 (49)

<sup>a</sup> The rms deviations between atoms in the refined native segment and atoms in the x-ray conformation are backbone, 0.20 Å; side chain, 0.18 Å; all, 0.19 Å. See also footnote a of Table III.

<sup>b-d</sup> See footnotes j–l of Table II.

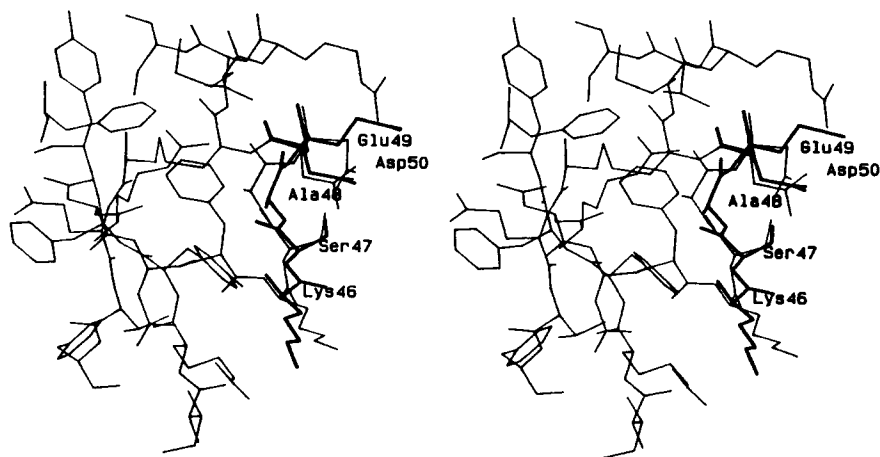
gave the lowest total energy does not have the best rms backbone atom deviation, only one of the best. For the conformation with the lowest total energy, the rms deviation of the backbone atoms from the x-ray structure is 2.00 Å before SA refinement, but the rms backbone atom deviation is 1.5 Å for a conformation 70 kcal/mol higher in energy. After 15 steps of MCSA refinement of the computed conformation with the lowest total energy, the total energy of the local region is higher than that for the refined native conformation by 79 kcal/mol, and the rms deviation of the backbone atoms for the refined computed segment from the native x-ray conformation is 1.87 Å. A problem associated with the removal of buried water molecules between the loop and the fixed part of the protein will be discussed below.

**Table VIII** Deviations of L9(10–18) from Native X-Ray Conformation in Ångströms<sup>a</sup>

	Backbone <sup>b</sup>	Side Chain <sup>c</sup>	All <sup>d</sup>
Tyr10	0.64 (5)	0.88 (7)	0.79 (12)
Thr11	1.71 (5)	2.61 (2)	2.01 (7)
Gly12	2.30 (4)	— (0)	2.30 (4)
Pro13	2.48 (6)	3.05 (1)	2.57 (7)
Cys14	1.96 (5)	1.82 (1)	1.94 (6)
Lys15	2.20 (5)	4.40 (4)	3.36 (9)
Ala16	2.11 (5)	— (0)	2.11 (5)
Arg17	1.80 (5)	3.00 (6)	2.52 (11)
Ile18	0.48 (5)	0.57 (3)	0.51 (8)
Total	1.87 (45)	2.61 (24)	2.16 (69)

<sup>a</sup> The rms deviations between atoms in the refined native segment and atoms in the x-ray conformation are backbone, 0.19 Å; side chain, 0.21 Å; all, 0.20 Å. See also footnote a of Table III.

<sup>b-d</sup> See footnotes j–l of Table II.



**Figure 3.** Local region for the computed H5(46–50) segment. The view is nearly the same as in Figure 1 after a 180° rotation in the plane of the page. See caption to Figure 2.

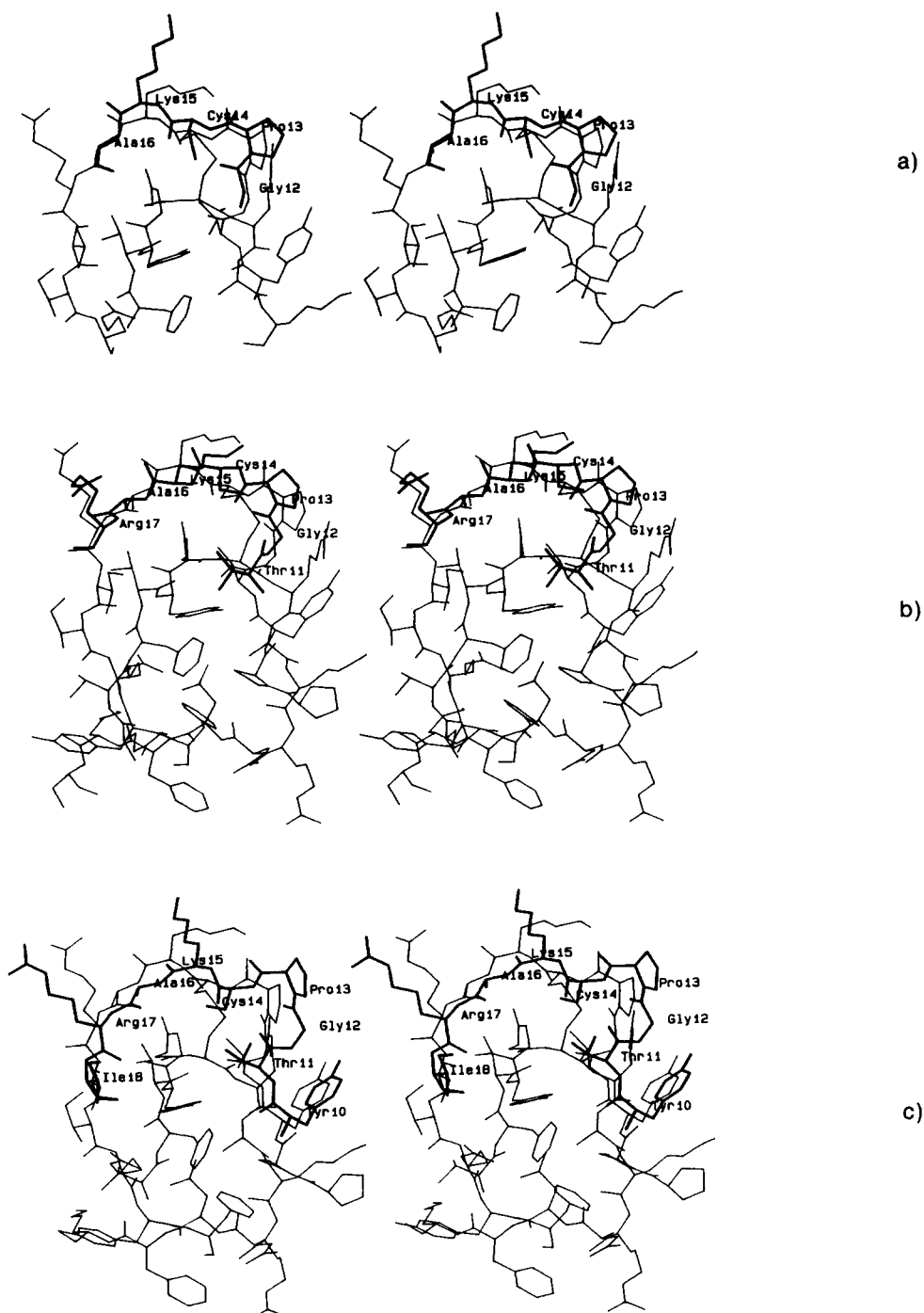
## DISCUSSION

An important aspect of the present approach to the calculation of loop conformations is the use of multiple limited simulations. Rather than run one long simulation that may spend a great deal of time far from the target conformation, multiple shorter simulations were performed. A comparison of the final total energies was used as a basis for ending the calculation and selecting the best result. The statistics for success in these trials seem reasonable. For example, in the calculation for B5(16–20) (Tables II and III), 4 simulations with 18K conformations per step and 4 with 24K conformations per step each produced one conformation close to the target native conformation. Also, the computed energies of these two conformations were the lowest, approaching the total energy of the refined native conformation.

For each calculation to compute a segment, except for L9(10–18), the conformation with the lowest energy had the smallest rms backbone atom deviation. For the short 5-residue segments, the rms backbone atom deviation is 1.1 Å or less. The energy discrepancy (difference between the refined native conformation and the predicted structure) for the  $\alpha$ -helical and the  $\beta$ -strands is 8 and 11 kcal/mol, respectively. This is due to breakage of a single peptide hydrogen bond (1.1 kcal/mol of stabilization based on one O—H interaction where the O—H distance is 1.9 Å<sup>25</sup>) plus the associated main-chain strain energy. The L5 segment discrepancy is about 2 kcal/mol. For the 7- and 9-residue loop segments, rms backbone atom deviations are 1.61 and 1.87 Å, respectively.

In the x-ray crystal structure there are three buried water molecules between the computed 9-residue loop and the fixed part of the protein. Hydrogen bonds connect the buried water molecules to the peptide hydrogens of Cys14 and Tyr10, and to the carbonyl oxygens of Thr11 and Tyr10. These were not included in the present calculations. Some of the distortion between the computed loop segments, L5(12–16), L7(11–17), and L9(10–18), and the native x-ray conformation of the segments, could have been due to further relaxation to fill in the voids left by these water molecules. Based on one O—H interaction in which the O—H separation is 1.7 Å, hydrogen bonds between water and peptide hydrogen and water and carbonyl oxygen provide 2.9 and 5.8 kcal/mol of stabilization, respectively. These energies were computed using the O and H energy parameters of the hydroxyl atom types for water.<sup>25</sup> We have considered the role that the three buried water molecules play in the calculation to refold the 9-residue loop. Results available indicate that inclusion of the three buried water molecules alone is not enough to lower the deviation between the predicted and the native loop conformations. The role of solvation may also be important.

Some of the approximations made may have affected the outcome in the calculation to compute H5(46–50). In these calculations, the neutral forms of all amino acids were used, and a constant dielectric of 2 was employed. The H5(46–50) segment has three ionizable side chains, Lys46, Glu49, and Asp50, and the local region around the segment has two more, Arg20 and Arg53, which may give rise to a number of interactions. The current methodology has some difficulties in modeling interactions on the protein surface. Simulations that include solvation



**Figure 4.** Local regions for the computed loop segments: (a) L5(12-16), (b) L7(11-17), and (c) L9(10-18). A disulfide bond links residues Cys14-Cys38. See caption to Figure 2.

energy are currently under way to better represent the solvation of polar atoms and the hydrophobic effect.

Alternative stabilizing interactions in the predicted conformation of H5(46-50) appear to have

disrupted the native hydrogen-bonding pattern. In the native x-ray structure, the Glu49 side chain on the outer surface of the protein is stabilized via hydrogen bonding to its own peptide hydrogen; in the predicted segment it is hydrogen bonded to the

Arg53 side chain. In the x-ray structure, the side chain of Arg53 appears to be hydrogen bonded to the side chain of Asp50, but in the predicted structure, the Asp50 side chain is hydrogen bonded to the Ser47 side chain. In the x-ray structure, the side chain of Ser47 appears to be hydrogen bonded to the peptide hydrogen of Asp50. Thus, the calculation shuffles hydrogen bonds in the x-ray structure that involve the surface-exposed side chains of Glu49, Asp50, Arg53, and Ser47. There is no reported solution structure for the surface side chains in BPTI; it seems likely that they sample more than one conformational state.

The rms overlap deviations in Table IV [computed according to Eq. (7)] indicate that the polypeptide continuity constraint [see discussion to Eqs. (4) and (6)] succeeded in smoothly connecting the computed segments with sequential residues on the fixed parts of the protein in the final structures. The rms overlap deviation reported in Table IV is less than 0.07 Å for the computed 5-residue segments and less than 0.14 Å for the computed 7- and 9-residue loop segments. For each residue on the fixed part of the protein sequentially connected to either the first or last residue of the computed segment, there are several internal coordinates associated with the shared peptide bond, namely the  $C'-N$  peptide bond length, the bond angles  $\angle C'-N-C^\alpha$ ,  $\angle O-C'-N$ ,  $\angle C'-N-C^\alpha$ , and  $\angle C'-N-H$ , and the improper torsion angles  $\angle N-O-C^\alpha-C'$  and  $\angle C'-H-C^\alpha-N$ . The rms overlap deviation reflects the departure of these internal coordinates from their standard geometries.

There appears to be an end effect in the reported rms deviations. The rms deviations of the computed segments from the native x-ray conformations, reported in Tables III and V–VIII, show that the smallest rms deviations occur for the first and last residues in the computed segment sequentially connected to a residue on the fixed part of the protein. The one exception is Arg17 in segment L7(11–17), which has a very large overall rms deviation. On the other hand, the main-chain conformation of this residue is very close to the native. The residues of the computed segment that are sequentially connected to the fixed part of the protein interact with (nest against) the fixed part of the protein much more than do the middle residues in the segments.

Recently, Lee and Subbiah<sup>15</sup> optimized the side-chain conformations of several proteins, whose structures are known (including BPTI), beginning from initially random conformations of the side chains, while the known main-chain conformations remained fixed. The MCSA protocol they used dif-

fered from ours in several ways. Among these, two are most noteworthy: only one long simulation with many more cooling steps was run, and only the van der Waals nonbonded energy was computed for the energy function in the annealing. In the present study, the entire segment, side chains and main chain, is computed, and all the loop segments computed are located on the surface of BPTI. In the calculation by Lee and Subbiah to optimize the side chains in BPTI,<sup>15</sup> the overall side-chain rms deviation was 2.61 Å, and the rms deviation for side chains of core residues was 1.65 Å, respectively. These calculated deviations included  $C^\beta$  atoms, whose position is held fixed along with the main chain. For the segments computed in this work, the overall rms deviation of side chains is 2.75 Å (excluding  $C^\beta$ , and  $C^\delta$  of proline). The rms deviation for side chains with temperature factors less than 25 is 1.86 Å. The major part of the rms deviation arises from the errors in just a few surface side chains. The rms deviations of surface side chains (characterized by large crystallographic temperature factors) and “core” side chains, reported in Tables III and V–VIII, are similar to the rms deviations reported by Lee and Subbiah.<sup>15</sup> It is our experience that the more interactions there are between the predicted side-chain atoms and the fixed atoms of the protein, the better is the prediction.

Table IX gives a summary of the computational effort required to compute the segments. Based on the number of conformations sampled per step in the simulations to predict the 5-residue  $\alpha$ -helical,  $\beta$ -strand, and loop segments, the computed  $\alpha$ -helical segment required the most effort (see row 5 in Table IX). Although the number of coordinates that are parameters in the annealing is nearly the same for the  $\alpha$ -helical segment and the  $\beta$ -strand (row 4 in Table IX), the helical segment required twice as many conformations. As Levinthal and co-workers point out, the more extended the chain segment, the fewer the acceptable conformations that can span this end to end distance.<sup>10</sup> Columns 3–6 in Table IX indicate that the computational effort required does not increase sharply with loop size, i.e., NCON increases from 18 to 30 to 42K as the loop size increases from 5 to 7 to 9 residues.

In comparing the efficiency of the MCSA approach, we used with the approach of Levinthal and co-workers<sup>2,10</sup> or the exhaustive search protocol employed by Brucoleri and Karplus,<sup>3,9</sup> one should keep in mind that the CPU time required to compute the correct loop conformation depends on the particular loop studied, the method, and the computer used in the calculation. In our study, the loops and the pro-

**Table IX Summary of Computational Effort for Each Segment Computed**

	H5(46-50)	B5(16-20)	L5(12-16)	L7(11-17)	L9(10-18)
Parameters <sup>a</sup>	30	34	21	34	43
NCON <sup>b</sup>	36K	18K	18K	30K	42K
rms <sup>c</sup>	0.94	1.11	1.03	1.61	1.87
Cray CPU <sup>d</sup>	3.0	1.8	1.2	3.4	6.2

<sup>a</sup> The total number of parameters for each simulation (dihedral plus rigid body coordinates).

<sup>b</sup> The number of conformations per step (1K = 1000) used to compute the lowest energy conformation.

<sup>c</sup> The rms deviation of backbone atoms (N, C $\alpha$ , C $\beta$ , C', O, and C $\delta$  of proline) between the computed and x-ray conformations of the segment.

<sup>d</sup> The CPU hours on a CRAY Y-MP/832 to run one simulation with the value of NCON indicated (not including refinement). (The CRAY Y-MP/832 is about 22 times faster than the SGI IRIS 4D25 processor for these calculations.)

tein are different from those studied by Bruccoleri and Karplus<sup>3</sup> and Levinthal and co-workers.<sup>2</sup> However, Bruccoleri and Karplus did compute 5-residue  $\beta$ -strand and helical segments. Our method appears to be more efficient than the method of Levinthal and co-workers. In order to obtain multiple hits on the lowest energy conformation, Levinthal and co-workers start from many initially screened loop conformations and alternate between energy minimization and molecular dynamics. Each one of these calculations requires several hundred hours of CPU on a Star ST-100 array processor, which is stated to be about 150 times faster than a MicroVAX II.<sup>2</sup> In the method of Bruccoleri and Karplus, program CONGEN, many conformations of the loop are generated by an exhaustive search of coordinate space. As they point out, the time it takes to do an exhaustive search increases exponentially with the number of degrees of freedom. The method we employ differs in that the computational effort increases less sharply with the loop size and is susceptible to great improvements in efficiency.

In work in progress, the MCSA approach employed here is being enhanced to make the calculation more efficient. The enhancements that are being tested include (a) varying the segment origin during the calculation, (b) varying the main-chain/side-chain selection probability, (c) altering the torsion angles in pairs, and (d) the use of rotamer libraries. With current modifications, the  $\beta$ -strand segment is predicted faster by a factor of 20 with no loss of hydrogen bonding, and the 9-residue loop is predicted faster by a factor of 3 and more accurately.<sup>40</sup>

## CONCLUSIONS

A MCSA approach was used to compute the conformation of five different loops in BPTI. In every

case but one (the conformation computed for a 9-residue loop segment), the conformation with the lowest total energy had the smallest rms deviation of the main-chain atoms when compared to the native x-ray conformation. The conformation of side chains having many interactions with the fixed part of the protein are predicted to the same level of accuracy as the main chain. The conformation of surface side chains are much less defined. The results obtained support the idea that it is more efficient to perform a number of independent limited simulations rather than to follow a single lengthy trajectory. This work demonstrates that the main-chain conformation of a 5-residue surface loop segment computed by the method developed here should deviate by little more than 1 Å from the experimental x-ray crystal conformation.

This research was supported in part by an NIH research grant and a grant from the Pittsburgh Supercomputing Center through the NIH Division of Research Resources Cooperative Agreement 1P41 RR06009-01 and through a grant from the National Science Foundation Cooperative Agreement ASC-8500650. We thank Profs. Jack S. Leigh, Harvey Rubin, and Peter Sterling for the use of their computer resources. We are grateful to Prof. Charles Brooks III for discussion and a careful reading of the manuscript.

## REFERENCES

1. Amzel, L. M. & Poljak, R. J. (1979) *Ann. Rev. Biochem.* **48**, 961-997.
2. Fine, R. M., Wang, H., Shenkin, P. S., Yarmush, D. L. & Levinthal, C. (1986) *Proteins* **1**, 342-362.
3. Bruccoleri, R. E., Haber, E. & Novotny, J. (1988) *Nature (London)* **335**, 564-568.
4. Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin,

- F., Beckmann, E. & Downing, K. H. (1990) *J. Mol. Biol.* **213**, 899–929.
5. Eriksson, A. E., Baase, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992) *Science* **255**, 178–183.
6. Sauer, R. T. & Lim, W. (1992) *Curr. Opin. Struct. Biol.* **2**, 46–51.
7. Englander, S. W. & Kallenbach, N. R. (1984) *Quart. Rev. Biophys.* **16**, 521–655.
8. Englander, S. W., Englander, J. J., McKinnie, R. E., Ackers, G. K., Turner, G. J., Westrick, J. A. & Gill, S. J. (1992) *Science* **256**, 1684–1687.
9. Brucoleri, R. E. & Karplus, M. (1987) *Biopolymers* **26**, 137–168.
10. Shenkin, P. S., Yarmush, D. L., Fine, R. M., Wang, H. & Levinthal, C. (1987) *Biopolymers* **26**, 2053–2085.
11. Chou, K. C., Némethy, G., Pottle, M. & Scheraga, H. A. (1989) *J. Mol. Biol.* **205**, 241–249.
12. Carlacci, L. & Chou, K.-C. (1990) *Biopolymers* **30**, 135–150.
13. Carlacci, L. & Chou, K.-C. (1990) *Protein Eng.* **3**, 509–514.
14. Chou, K. C. & Carlacci, L. (1991) *Proteins* **9**, 280–295.
15. Carlacci, L., Chou, K.-C. & Maggiora, G. M. (1991) *Biochemistry* **30**, 4389–4398.
16. Lee, C. & Subbiah, S. (1991) *J. Mol. Biol.* **217**, 373–388.
17. Lee, C. & Levitt, M. (1991) *Nature (London)* **352**, 448–451.
18. Wilson, S. R. & Cui, W. (1990) *Biopolymers* **29**, 225–235.
19. Brünger, A. T. (1990) *Molecular Dynamics: Applications in Molecular Biology*, CRC Press Topics in Molecular and Structural Biology series, Goodfellow, J. M., Ed., CRC Press, Boca Raton, FL, pp. 137–178.
20. Brünger, A. T., Kuriyan, J. & Karplus, M. (1987) *Science* **235**, 458–460.
21. Wüthrich, K. (1989) *Acc. Chem. Res.* **22**, 36–44.
22. Clore, G. M. & Gronenborn, A. M. (1989) *CRC Crit. Rev. Biochem.* **24**, 479–564.
23. Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1979) *J. Phys. Chem.* **79**, 2361–2381.
24. Chou, K. C., Némethy, G., Rumsey, S., Tuttle, R. W. & Scheraga, H. A. (1985) *J. Mol. Biol.* **186**, 591–609.
25. Némethy, G., Pottle, M. S. & Scheraga, H. A. (1983) *J. Phys. Chem.* **87**, 1883–1887.
26. Carlacci, L. & Chou, K.-C. (1990) *Protein Eng.* **4**, 225–227.
27. Go, N. & Scheraga, H. A. (1970) *Macromolecules* **3**, 178–187.
28. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. J. (1953) *Chem. Phys.* **21**, 1087–1092.
29. Kirkpatrick, S., Gelatt, C. D., Jr. & Vecchi, M. P. (1983) *Science* **220**, 671–680.
30. Whille, L. T. (1986) *Nature (London)* **324**, 46–48.
31. Wilson, C. & Doniach, S. (1989) *Proteins* **6**, 193–209.
32. Chou, K. C. & Carlacci, L. (1990) *Protein Eng.* **4**, 661–667.
33. IUPAC-IUB Commission on Biochemical Nomenclature. (1970) *J. Mol. Biol.* **52**, 1–17.
34. Marquart, M., Walter, J., Deisenhofer, J., Bode, W. & Huber, R. (1983) *Acta Crystallogr. Sect. B* **39**, 480.
35. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112**, 535–542.
36. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Protein Data Bank in Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, Allen, F. H., Bergerhoff, G. & Sievers, R., Eds., Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.
37. Gay, D. M. (1983) *Assoc. Comput. Mach. Trans. Math. Software* **9**, 503–524.
38. Vasquez, M., Némethy, G. & Scheraga, H. A. (1983) *Macromolecules* **16**, 1043–1049.
39. Wagner, G. & Wüthrich, K. (1982) *J. Mol. Biol.* **160**, 343–361.
40. Carlacci, L. & Englander, S. W., in preparation.

Received September 21, 1992

Accepted February 3, 1993