A Genetic Algorithm that Seeks Native States of Peptides and Proteins

Shaojian Sun
Structural Biochemistry Program, Frederick Biomedical Supercomputing Center, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, Maryland 21702 USA

ABSTRACT We describe a computer algorithm to predict native structures of proteins and peptides from their primary sequences, their known native radii of gyration, and their known disulfide bonding patterns, starting from random conformations. Proteins are represented as simplified real-space main chains with single-bead side chains. Nonlocal interactions are taken from structural database-derived statistical potentials, as in an earlier treatment. Local interactions are taken from simulations of (ϕ,ψ) energy surfaces for each amino acid generated using the Biosym Discover program. Conformational searching is done by a genetic algorithm-based method. Reasonable structures are obtained for melittin (a 26-mer), avian pancreatic polypeptide inhibitor (a 36-mer), crambin (a 46-mer), apamin (an 18-mer), tachyplesin (a 17-mer), C-peptide of ribonuclease A (a 13-mer), and four different designed helical peptides. A hydrogen bond interaction was tested and found to be generally unnecessary for helical peptides, but it helps fold some sheet regions in these structures. For the few longer chains we tested, the method appears not to converge. In those cases, it appears to recover native-like secondary structures, but gets incorrect tertiary folds.

INTRODUCTION

Computational protocol that can fold a protein to its native conformation from its primary sequence information alone has not been successfully developed. It has been realized that simplified models of protein structure and interactions are needed to overcome the enormous complexicity of this protein folding problem, which in part is due to many degrees of freedom of a protein structure, complicated interactions among amino acids and their environment, and multiple minima in the free energy landscape for such a system. Great effort has been devoted to develop different simplified models and conformational search strategies in the past decades (Levitt and Warshel, 1975; Levitt, 1976; Kuntz et al., 1976; Hagler and Honig, 1978; Tanaka and Scheraga, 1976; Crippen and Viswanadhan, 1984, 1985; Wilson and Doniach, 1989; Skolnick and Kolinski, 1989, 1991; Kolinski and Skolnick, 1994; Covell and Jernigan, 1990; Fiebig and Dill, 1993; Yue and Dill, 1995; Sippl et al., 1992; Hinds and Levitt, 1992; Unger and Moult, 1993; Bowie and Eisenberg, 1994; Dandekar and Argos, 1992, 1994; Sun et al., 1992; Sun, 1993, 1995), and the successful structural predictions are limited to small proteins that contain most helical and/or strand structures.

We recently developed a computational procedure that attempts to find the three-dimensional native structures of proteins and peptides given the amino acid sequence, the native radius of gyration, and the disulfide bonding pattern of the protein (Sun, 1993). It used a real-space but simplified representation of amino acid chains, a genetic search algorithm, and a simple potential function. In that work, the potential function consisted of two nonlocal interaction terms and two local interaction terms. The nonlocal interaction was a statistical potential derived from a database of known protein structures. The two local terms were also derived from populations observed in a structural database: one term described the torsion angle preferences of each amino acid monomer; the other term described dipeptide pair conformational preferences.

Here we use the same chain representation and search strategy and show that predictions can be both improved and given a firmer physical basis by now using a much simpler local interaction term. Now for local interactions, we simply use (ϕ,ψ) energy maps that were generated using the Biosym (San Diego, CA) torsional forcing method and the Discover force field. We have generated all 20 such maps for each amino acid. The nonlocal interactions remain represented by a database-derived statistical potential function. Here we perform the conformation searches for melittin, apamin, avian polypeptide inhibitor, crambin, and some small designed peptides. We find the results to be sensibly consistent with the known crystal structures, circular dichroism (CD), and nuclear magnetic resonance (NMR) measurements. We also discuss a few larger molecules where the method fails to generate native tertiary structures, although in those cases it usually does fairly well in predicting the α-helix and β-strand elements.

SIMPLIFIED CHAIN REPRESENTATION

The protein chain is represented as follows (Sun, 1993; Wilson and Doniach; 1989, Sun, et al., 1992) (Fig. 1). 1) All backbone bond lengths and bond angles have their ideal values (Corey and Pauling, 1953). 2) All the peptide bond dihedral angles are fixed in the *trans* (ω = 180°) confor-

Received for publication 19 January 1995 and in final form 18 April 1995.

Address reprint requests to Dr. Shaojian Sun, Structural Biochemistry Program, Frederick Biomedical Supercomputing Center, SAIC, National Cancer Institute, Frederick Cancer Research and Development Center, Frederick, MD. Tel.: 301-846-1989; Fax: 301-846-6066; E-mail: sun@daedalus.ncifcrf.gov.

0006-3495/95/08/340/16 $2.00

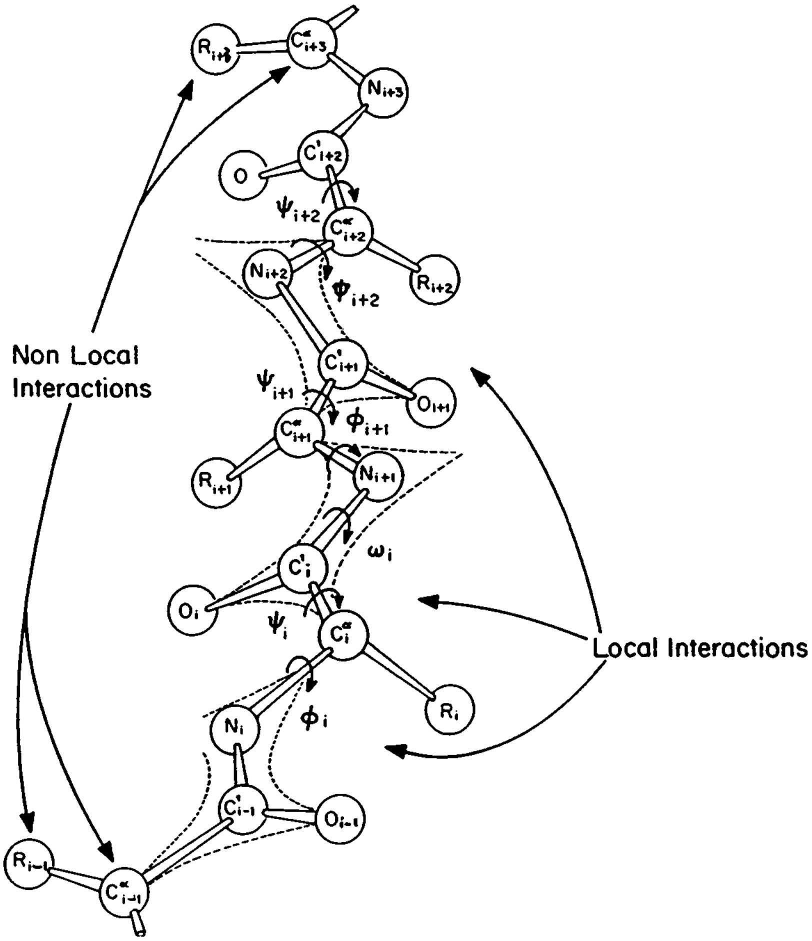

FIGURE 1 Current reduced representation model of protein structure (from Sun, 1993). Full backbone and side chain centroids have been preserved in the model bond lengths, and bond angles are held to their ideal values, dihedral angle $\omega = 180°$, and the geometric variables are ϕ, ψ at both sides of C^{α}. A side chain is represented by a point located at the average position of the side chain-heavy atoms. A statistical nonlocal potential function and a physical averaged local potential function are used as the objective functions of the system in the genetic algorithm conformation minimization.

mation. 3) A single spherical virtual atom is used to represent each side chain at the center of mass of the heavy atoms in the side chain. 4) Conformational freedom is represented by the backbone dihedral angles ϕ and ψ.

The (ϕ,ψ) internal coordinate representation has the advantages 1) of fewer variables than cartesian coordinates, 2) of simplifying the neglect of covalent bond stretching that allows us to eliminate a large background energy that otherwise obscures other more important conformational energies, and 3) of simplifying the use of Ramachandran plot information about torsional preferences of amino acids.

We use both the root mean square (RMS) and the distance matrix (DME) errors to compare differences among structures. The RMS and DME values are computed for the backbone atoms. They are defined by

$$RMS = \left\{\frac{1}{N}\sum_{i}^{N}(r_i - r_i^c)^2\right\}^{1/2} \qquad (1)$$

$$DME = \left\{\frac{2}{N(N-1)}\sum_{ij}^{N}(r_{ij} - r_{ij}^c)^2\right\}^{1/2} \qquad (2)$$

where superscript c indicates a reference conformation, which we usually take to be the known native structure from crystallography or NMR. We also consider two other measures of similarity, which are radius of gyration and the total number of contacts among all the residues. A contact arises when the C^{α} carbons of two residues are separated less than 10.0 Å.

POTENTIAL FUNCTION

Fig. 1 defines the local and nonlocal interactions in the model. Local interactions are those along the peptide bond among nearest neighboring amino acids. They are functions of the backbone dihedral angles (ϕ,ψ). Nonlocal interactions occur among monomers separated in the primary sequence by at least two intervening residues. The nonlocal interaction depends on the spatial separations between two residues 1) of the C^{α} carbons, and 2) between the centroids of the side chains. The total energy is:

$$H = E_{\text{local}}(\{\phi_i, \psi_i\}) + E_{\text{nonlocal}}(\{r_{ij}^{C^{\alpha}}, r_{ij}^{SC}\}) \qquad (3)$$
$$= \gamma_1 \sum_{i} E_{k_i}^{S}(\phi_i, \psi_i) + \gamma_2 \sum_{i<j-2} E_{k_ik_j}^{C^{\alpha}}(r_{ij}^{C^{\alpha}}) + \gamma_3 \sum_{i<j-2} E_{k_ik_j}^{SC}(r_{ij}^{SC})$$

where i or j is the sequence position of a residue, k_i is the amino acid type of residue i, $r_{ij}^{C^{\alpha}}$ is the distance between the C^{α} atoms of the residues i and j, r_{ij}^{SC} is the distance between the side chain centroids of the residues i and j, $E^{\mathrm{S}}_{\mathrm{k_i}}$ is the singlet local interaction potential, $E^{\mathrm{C}^\alpha}_{\mathrm{k_ik_j}}$ is the C^α interaction potential, and $E^{\mathrm{SC}}_{\mathrm{k_ik_j}}$ is the side chain centroid interaction potential. γs are empirical coefficients that represent the relative importance of the various terms in the potential energy function.

This model-potential function Eq. 3 differs from the one in our previous studies (Sun, 1993; Sun et al., 1992) in two aspects. First, it uses a physical local potential function to replace the statistical local potential function, and secondly, it does not include the local dipeptide interaction term. Computing a dipeptide term is much more involved, and we have no evidence that the added complexity is warranted.

MODELING THE LOCAL INTERACTIONS

In the present work we use the Consistent Valence Force Field (CVFF) force field from Biosym to generate the torsional energy landscape, $E\,(\phi,\psi)$ for each of the 20 amino acids. In these simulations, we have not included solvent. Each such energy is computed by the method of torsional forcing (Stern et al., 1983). The torsional-forcing strategy systematically biases the torsion angle toward each given ϕ and ψ value by adding a penalty function to the total molecular energy. The penalty function is harmonic in the deviation of the forced angle from its target value, i.e.,

$$\mathit{Total\ energy} = V(\phi\,,\psi) + K(\phi - \phi^0)^2 + K(\psi - \psi^0)^2 \quad (4)$$

where $V(\phi,\psi)$ is the energy of the all-atom representation of the molecule, (ϕ^0,ψ^0) are the desired target dihedral angles, and K is the force constant that scales the strength of the biasing function. By minimizing the total energy over all degrees of freedom including the penalty weights, the structure can respond to the forced torsion. Typically, the final structure is lower in energy than if only the (ϕ,ψ) angles had been allowed to change and is therefore a more realistic representation of the true energy surface.

The force field we use is the CVFF in the Discover program of Biosym:

$$\begin{aligned}
E_{\mathrm{pot}} = {} & \sum_{\mathrm{b}} D_{\mathrm{b}}(1 - e^{-\alpha(\mathrm{b}-\mathrm{b_0})})^2 + 1/2 \sum_{\theta} H_\theta(\theta - \theta_0)^2 \\
& + 1/2 \sum_{\phi} H_\phi[1 + s \cdot \cos(n\phi)] + 1/2 \sum_{\chi} H_\chi \chi^2 \\
& + \sum_{\mathrm{b}} \sum_{\mathrm{b'}} F_{\mathrm{bb'}}(b - b_0)(b' - b_0') \\
& + \sum_{\phi} \sum_{\phi'} F_{\theta\theta'}(\theta - \theta_0)(\theta' - \theta_0') \qquad (5) \\
& + \sum_{\mathrm{b}} \sum_{\theta} F_{\mathrm{b}\theta}(b - b_0)(\theta - \theta_0) \\
& + \sum_{\phi} F_{\phi\theta\theta'}(\theta - \theta_0)(\theta' - \theta_0') + \sum_{\chi} \sum_{\chi'} F_{\chi\chi'}\chi\chi' \\
& + \sum \epsilon[(r^*/r)^{12} - 2(r^*/r)^6] + \sum q_{\mathrm{i}}q_{\mathrm{j}}/\epsilon r_{\mathrm{ij}}
\end{aligned}$$

The first four terms are the diagonal terms of the valence force field and represent the energy of deformation of bond lengths, bond angles, torsional angles, and out-of-plane interactions, respectively. The Morse potential (term 1) is used for bond stretching. Terms 5–9 are off-diagonal terms and represent couplings between deformations of internal coordinates. Terms 10 and 11 describe the nonbonded interactions. Term 10 represents the van der Waals interaction with a Lennard-Jones function. Term 11 is the Coulombic representation of electrostatic interactions.

We set the parameters as follows: the forcing constant in equation (4) $K = 500.0$ kcal; and a forcing constant of 100.0 kcal is used to hold the peptide bond planar, $\omega_1 = \omega_2 = 180°$ (*trans* conformation). The dielectric constant equals one in our calculations. The (ϕ,ψ) energy surface is computed on a 36 × 36 grid (10° interval for both ϕ and ψ). Each point on the grid in (ϕ,ψ) is first optimized by a steepest descent minimization method to have a maximum derivative less than 1.0 kcal/Å, and is then further optimized by a quasi-Newton-Raphson minimization procedure (Fletcher, 1972) so that the maximum derivative is less than 0.05 kcal/Å.

The (ϕ,ψ) map is then further refined by interpolating to get a grid of 72 × 72 (5° intervals for both ϕ and ψ). Interpolation reduces computer time by 75%. On average, a dipeptide (ϕ,ψ) energy surface calculation takes about 2.0 CPU h on an SGI personal Iris for a 36 × 36 grid energy surface.

The energy scaling coefficients (γ_1, γ_2, γ_3) between the local and the nonlocal terms have been set to 1.0, 1.0, and 1.0, respectively, as before (Sun, 1993). In comparison with the database-derived local potentials we used before, the present energies vary over a larger range. For the two nonlocal terms in Eq. 3, we use a cut off distance of 15 Å; beyond that, pair energies equal zero.

CONFORMATIONAL SEARCH ALGORITHM

We use a genetic algorithm (Holland, 1975; Goldberg, 1989) for conformational searching. The details of our algorithm are described in (Sun, 1993); other implementations of genetic algorithms to protein folding are described in (Tuffrey et al., 1991; Blommers et al., 1954; Dandekar and Argos, 1992, 1994; Bowie and Eisenberg, 1994). Genetic algorithms are patterned after natural selection and genetic processes (Holland, 1975; Goldberg, 1989). Our method works as follows.

A structure of a protein having a given primary sequence is represented by a sequence of $\binom{\phi}{\psi}$ values, called the conformational string. The genetic algorithm search involves three steps:

1) Set up the initial population of conformations. Randomly create a population of $\binom{\phi}{\psi}$ conformational strings. Each string represents one specific conformation. Compute the conformational energy for each string.

2) Genetic operations–replication, mutation and crossover–are used to change certain $\binom{\phi}{\psi}$ values in the population

of the conformation strings. The mutation operation randomly chooses one or several $\binom{\phi}{\psi}$ sites and changes their values randomly. Replication generates a population of such mutated conformational strings. The crossover operation randomly selects two conformational strings and cuts them at a random site, then generates two new strings by interchanging the parts from the old strings. Replication is also used here to generate a population of crossed-over strings. The strings from a previous generation are copied to the next generation by replication.

3) Select the new generation of conformational strings. Compute the conformational energy for each string in replication, mutation, and crossover populations, and select a new population of strings that have lowest energies. To further facilitate the conformational search, we used a segmental mutation method in which a group of $\binom{\phi}{\psi}$ changes simultaneously according to the Ramachandran distribution (Sun, 1993).

The number of starting conformations is 90 for the melittin simulation, so the mutation population is 180. We start with 200 conformations for crambin and the other small proteins, so the mutation population size is 400. A random monogamy crossover is performed among different conformational species in each generation to create a crossover population. The crossover population is created between the last generation and the newly created mutation conformation population, and has a size that is twice as large as the initial population (180 for melittin and 400 for others). The monogamy crossover is defined that one conformation species can only crossover with one and only one other conformation species. Initial conformations were created randomly from the primary segmental conformation pools with an additional random perturbation on each (ϕ,ψ) angle.

The overall segmentation probability (see Appendix) is $(P_2, P_3, P_4, P_5) = (0.4, 0.3, 0.2, 0.1)$, except when noted otherwise. This choice is based on the fact that the larger the probability for the shorter segments, the higher the variability in the constructed conformations. The same segmentation probabilities have been used in the mutation operation in which both of the segmental lengths and the mutation sites in a conformational species are randomly chosen. We have chosen 1 as the number of simultaneous mutation sites for the mutation operation in a conformational species. The partition of the segmentation for any conformational species in different generations is uncorrelated; in other words, we repeat the random segmentation for all conformational species in every generation. The search process is terminated when no lower energy conformations are found in 20 consecutive generations of the search.

EXTRA CONSTRAINTS

The conformational space of most proteins is too vast to be explored efficiently using our present combination of simplified chain representation, genetic search algorithm, and potential function. Hence, we limit the present study to peptides and only the smallest proteins, and we apply additional constraints. For learning the limitations of potential functions and search strategies, constraints are helpful because if a method does not succeed under appropriate constraints, it will surely not succeed without them. In some of the cases tested below, we use an additional energy term to bias the conformational search toward the native radius of gyration:

$$E_{rg} = \lambda(R_g - R_g^{\text{native}})^2 \qquad (6)$$

where

$$R_g = \sqrt{(1/N) \sum_{k=1}^{N} (\vec{C^\alpha} - \vec{CM})^2} \qquad (7)$$

λ is a penalty coefficient, and $\vec{CM}$ are the center of mass coordinates. We previously found that chains can often fold to near native compactness without this constraint (Sun, 1995); however, this term increases the efficiency of the conformational search.

Second, we introduce known disulfide bond constraints for crambin and tachyplysin-I. For disulfide bonds, we assume an energy

$$E_{ss} = \sum_{i=1}^{m} \lambda_{ss}(D_{C^{sc}-C^{sc}} - d_s)^2 \qquad (8)$$

where λ_{ss} is a penalty coefficient, $D_{C^{sc}-C^{sc}}$ is the distance between two cysteine side chain centroids that form the disulfide bridge, d_s is the optimal $C^{sc} - C^{sc}$ distance for a disulfide bond. We use $d_s = 2.71$ Å and we found $\lambda_{ss} = 20.0$ kcal/Å to work well.

In some cases, which are mentioned explicitly, we have also included a model hydrogen bond interaction term. A hydrogen bond is a 5.0 kcal/mol attraction when the distance between main-chain hydrogen bonding atoms O and H is <2.5 Å, and the N–H–O angle is 120–180°.

RESULTS

We report here the results of simulations on melittin, a membrane protein of 26 residues; 36-residue avian pancreatic polypeptide inhibitor (APPI); crambin, a protein of 46 residues with three disulfide bonds; apamin, an 18-residue polypeptide component of bee venom containing two disulfide bonds; tachyplesin, a 17-mer antimicrobial cationic polypeptide with two disulfide bonds; C-peptide, the first 13-residue segment of ribonuclease A, two short α-helix sequences designed by Marqusee and Baldwin (1987); a short α-helix designed by Hill et al. (1990); and a 26-mer α-helix sequence designed by Klaus and Moser (1992). We also compute the α-helix formation probability of the N-terminus fragment of Barnase, which has been studied experimentally by CD and

NMR by Fersht and colleagues (Sancho et al., 1992; Fersht et al., 1992). We discuss results for a zinc finger protein, ubiquitin, and cytochrome 256, where the method failed to generate the native structures.

Melittin

We computed the structure of melittin and compared it to the crystal structure, which has a resolution of 2.0 Å (Terwilliger and Eisenberg 1982). The results, shown in Table 1, were based on using the radius of gyration as a constraint, with penalty coefficient λ in Eq. 6, set to 200 kcal/Å. The hydrogen bonding energy term was not used. The random perturbation *d* on (ϕ,ψ) angles was chosen to be 10°. 90 structures have been optimized simultaneously. The initial total energy E_{start}, of each of the 90 initial structures ranges widely from 471.32 to 14,628.35 kcal. The average starting energy of 90 structures is 1856.04 kcal with a standard deviation of 1952.37 kcal. The total energy at the end of the simulation converges to a mean value of 358.39 kcal with a standard deviation of 0.30 kcal/kcal units (Table 1). The average DME and RMS errors of the 90 computed structures to the crystal structure are 0.66 Å and 0.85 Å, respectively. The number of contacts averaged over all 90 optimized structures is 307.4 with a standard deviation of 1.7. The average radius of gyration is 10.8 Å. These values correspond closely to the crystal structure, which has 300 total contacts and a radius of gyration of 11.1 Å. The computed structures converge uniformly not only in their final energies, but also in their three-dimensional conformations. The same has been found in our previous study (Sun, 1993). There is high degree of similarity among the 90 computed structures. The RMS error between any two of the 90 computed melittin structures is <0.20 Å (Fig. 2 *e*). Fig. 2 shows a stereo plot of the backbone and side chain centroid for the melittin crystal structure (Fig. 2 *a*) and one of the computed structures (Fig. 2 *b*). In all the computed structures, there is a bending at the middle of the structure due to proline 14, consistent with the melittin crystal structure.

The present model gives improved structures compared to the earlier model. Fig. 2 *c* is one of the computed melittin structures by the all-statistical potential. The folding around the N-terminal end in the melittin structures by the statistical potential is not as helical as that shown in the crystal structure of melittin. The physical potential does better in that regard, and it leads to smaller DME and RMS errors in comparison with the crystal structure.

When the explicit hydrogen bonding energy is added, the computed structures become slightly worse: they have an average DME of 1.0 Å, an average RMS 1.7 Å, and an average number of total contacts of 298.4. Fig. 2 *d* shows one of the computed structures with hydrogen bonding.

If we use the (ϕ,ψ) angles taken directly from the melittin crystal structure to compute a melittin structure in the current model, the resultant backbone RMS error is ~1.5 Å compared to the original crystal structure.

Apamin

Apamin is an 18-residue polypeptide component of bee venom containing disulfide bonds between residues 1 and 11 and between 3 and 15. No x-ray crystal structure is available, but there are NMR structures (Wemmer and Kallenbach, 1983). To establish the "experimental" native structure, we used the apamin backbone dihedral angle data from Freeman et al. (1986) and energy-minimized it using the Discover force field.

In the simulations for apamin, the segmentation probabilities have been set to $P_2 = 0.6, P_3 = 0.4, P_4 = P_5 = 0.0$ (see Appendix). Random perturbation of (ϕ,ψ) was used with $d = 10°$. Input to the algorithm was the primary sequence and disulfide bond constraints. We included the hydrogen bonding term in this case and the disulfide bond penalty energy term but not the radius of gyration constraint.

Starting from randomly created initial conformation population, 200 structures are computed. Table 2 lists the average properties of these computed structures. As expected, the initial energy profiles are high and the standard deviation is large. The energy profiles for the optimized structures are uniform and have a small standard deviation. The α-helical region of the C-terminal part appeared naturally as the result of the conformational search. The computed structures have an average total

TABLE 1 Simulations for melittin, a protein of 26 residues

Parameters	E_{start}	E_{end}	$E_{C\alpha}$	E_{SC}	E^{p}_{S}	DME (Å)	RMS (Å)	Total contacts	Radius of gyration (Å)
Average	1856.04	358.39	306.04	317.77	−265.87	0.66	0.85	307.4	10.8
σ	1952.37	0.30	0.48	0.57	0.23	0.02	0.02	1.7	0.0
Crystal Structure		418.09	324.82	339.79	−246.52			300.0	11.1

90 structures have been computed simultaneously. E_{start} is the energy of starting conformations, E_{end} the energy of the structures after the RRM genetic algorithm minimization, and E_{CS}, E_{SC}, and E^{p}_{S} are the energy components of C^{α}-C^{α} interaction, sidechain-sidechain interaction, and the local physical averaged singlet interaction, respectively, of the minimized structures. No explicit hydrogen bonding term is used. DME and RMS (unit in Å) are computed by using the crystal structure as the reference. σ denotes the standard deviation. The penalty coefficient λ in the E_{rg} has been set to 200 kcal/Å. The starting conformations were randomly created.

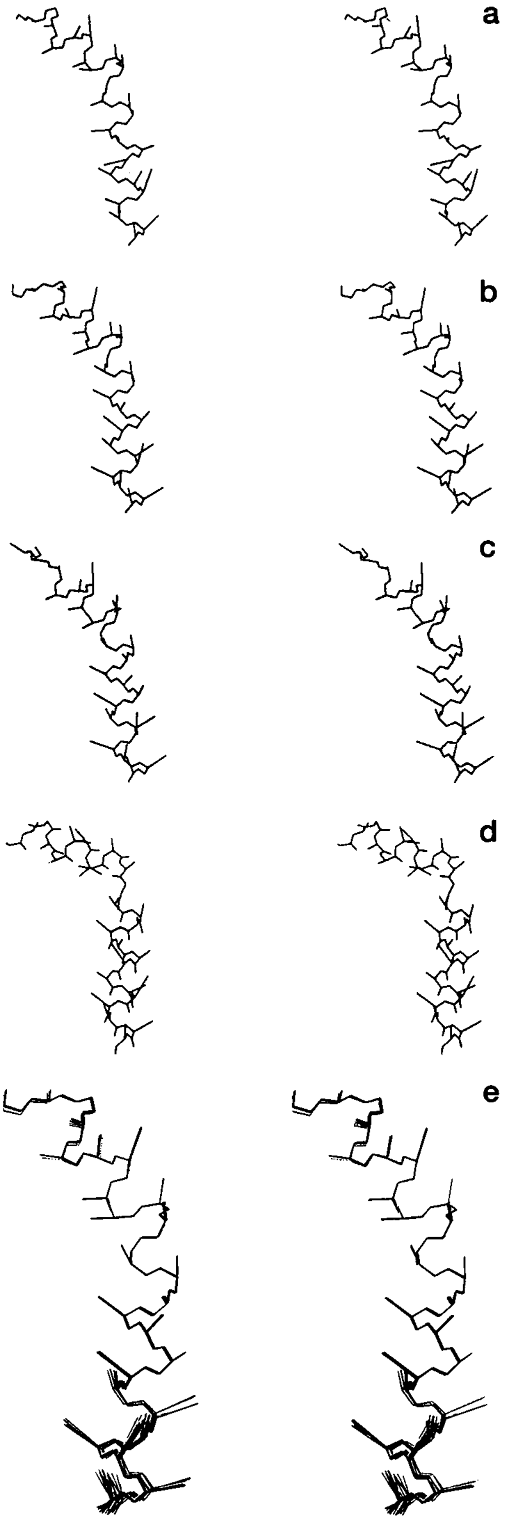

FIGURE 2 Backbone and side chain centroid stereo plot for melittin. (*a*) Crystal structure. (*b*) One of the 90 computed structures (*RMS* = 0.85 Å, *DME* = 0.65 Å to the crystal structure). (*c*) One of the computed structures generated by using the statistical potential of both local and nonlocal interaction (*RMS* = 1.64 Å, *DME* = 0.98 Å to the crystal structure). (*d*) One of the computed structures generated by using the physical potential and the explicit hydrogen bonding energy term (*RMS* = 1.7 Å, *DME* = 1.0 Å to the crystal structure). (*e*) Superposition of 30 computed melittin structures; these are all very similar and have an *RMS* <0.20 Å between any two structures.

contacts of 252 and an average radius of gyration of 5.95 Å, which are close to the corresponding values of 238 and 6.4 Å for the apamin NMR structure. In comparison with our previous results on apamin (Table 5, case b in Sun, 1993), the computed structures are substantially improved in terms of both DME and RMS. The computed structures are quite similar to each other. The RMS between any two computed structures is <1.0 Å (most of them are <0.5 Å).

Fig. 3 plots the DISCOVER-minimized apamin NMR structure (Fig. 3 *a*), and one of the computed apamin structures (Fig. 3 *b*). It was found experimentally (Pease et al., 1990) that if the nine residues at the apamin C-terminal half are replaced with those from the S-peptide from ribonuclease-A, this hybrid sequence folds to apamin like conformations. The S-peptide, consisting first 20 amino acids from the ribonuclease A has been shown to have a helix-forming propensity (Kim et al., 1982). As a test, we have computed the structure of the hybrid, using the same simulation conditions.

The apamin sequence and the hybrid sequence are listed below:

C-N-C-K-A-P-E-T-A-L-C-A-R-R-C-Q-Q-H Apamin sequence

C-N-C-K-A-P-E-T-A-A-C-K-F-E-C-Q-H-M Hybrid sequence

It would be desirable to carry out a computational experiment of the same kind. All the simulation conditions have been kept the same as those in the apamin simulation.

Table 3 summarizes the results. All the computed structures for the hybrid apamin sequence are like native apamin, the average DME and RMS values for 200 computed structures being 1.70 Å and 2.17 Å relative to the NMR apamin structure. Energies and three-dimensional structures converge. Fig. 3 *c* is a stereo plot of one of the computed structures for the hybrid apamin sequence. The C-terminal end forms naturally an α-helix as a result of the energy minimization, in agreement with the experimental data.

Tachyplesin-I

Tachyplesin-I is a cationic peptide of 17 residues with two disulfide bonds. Its structure is probably fairly rigid because of the restriction imposed by the two disulfide bonds at (3–16 and 7–12). It has no x-ray structure but an NMR study by Kawano et al. (1990) indicates that it forms an anti-parallel β-sheet-like structure.

We computed the structure of tachyplesin-I. We input the primary sequence and the disulfide bonds. Explicit hydrogen bonding term and disulfide bonding terms were used in the simulation.

Fig. 4 shows a backbone and side chain centroid stereo plot for one of the computed structures of tachyplesin-I. It shows an anti-parallel β-sheet-like conformation, consistent with the NMR experiment. We found that residues 8–11

TABLE 2 Simulations for apamin, a polypeptide of 18 residues with two disulfide bonds

Parameters	E_{start}	E_{end}	$E_{C\alpha}$	E_{SC}	E_S^p	DME	RMS	Total contacts	Radius of gyration
Average	10301.49	−115.40	104.99	150.59	−389.49	1.83	2.14	252.3	5.95
σ	6411.28	3.65	3.74	1.81	1.79	0.04	0.06	1.6	0.07
Crystal Structure		−66.72	158.24	174.38	−398.43			238	6.4

200 structures have been computed simultaneously. The reference structure used in the calculation of DME and RMS of the optimized structures was a DISCOVER-minimized NMR structure of apamin. σ denotes the standard deviation. Radius of gyration constraint is not used in the simulation. An explicit hydrogen bonding and disulfide bridge interaction energy terms are used.

formed a β-turn between strands 3–8 and 10–14, although Fig. 5 shows that the distance between residues 4–6 and residues 13–15 is too far to form hydrogen bonding. Because no experimental structure is available, these are untested predictions.

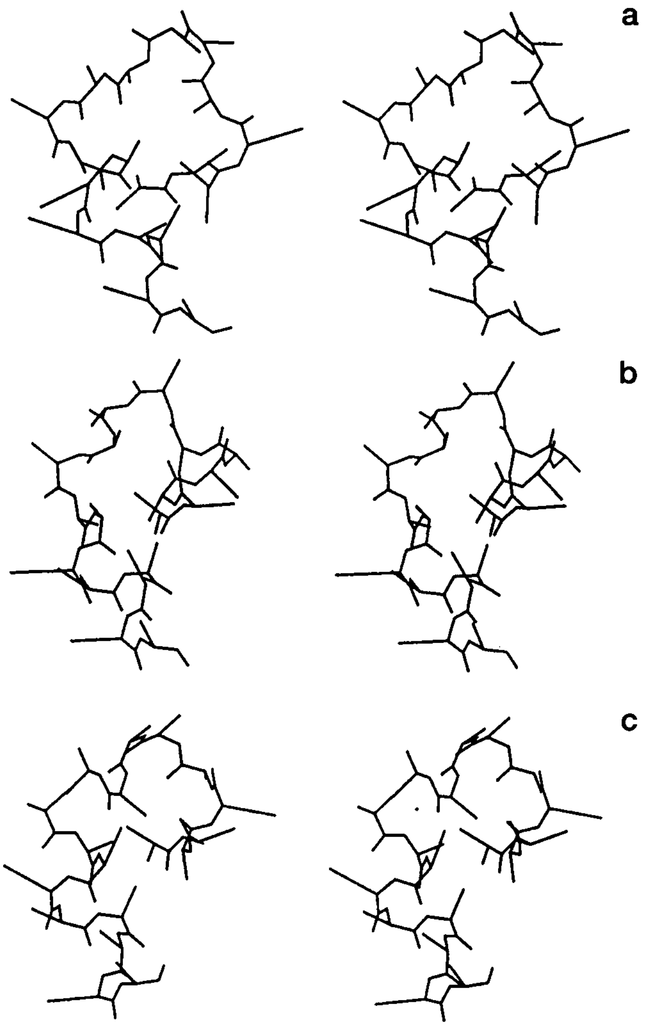

FIGURE 3 Backbone and side chain centroid stereo plot for apamin. (*a*) DISCOVER-minimized apamin NMR structure. (*b*) One of the computed structures (*RMS* = 2.13 Å, *DME* = 1.83 Å to the DISCOVER minimized apamin NMR structure). (*c*) One of the computed structures for the hybrid apamin sequence (*RMS* = 2.12 Å, *DME* = 1.65 Å).

APPI

APPI is a small protein of 36 residues, the structure of which is known at a resolution of 1.37 Å (Blundell et al., 1981; Glover et al., 1983). Inputs to the algorithm were only the primary sequence and the radius of gyration from the crystal structure, with penalty coefficient λ set to 200 kcal/Å. The size of the conformation population was set to 90. Table 4 summarizes the results for the computed structures. The computed structures have an average number of total contacts of 448, which is much less than the value of 530. The computed structures are less compact than the crystal structure (Fig. 5) and have large DME and RMS errors to the crystal structure. We also found that the computed structures had a much lower energy than the crystal structures. The crystal structure has higher nonlocal interaction energy than the computed structures. This indicates the inadequacy of the statistical nonlocal potential function used in this study, and it should be further improved. Nevertheless, the computed structures have correct α-helix motif found in the APPI crystal structure. The computed structures have α-helices from residues 14–28, while the crystal structure has a α-helix from residues 14–31.

Crambin

The crambin crystal structure has a resolution of 1.5 Å (Hendrickson and Teeter, 1981). For crambin simulation, we use the amino acid sequence, native radius of gyration λ = 200 kcal/Å), and the three known pairs of disulfide bonds (between 3 and 40, 4 and 32, and 16 and 26).

The disulfide bond potential function, Eq. 8, is applied in the simulation with the following parameters: the penalty coefficient, λ_{ss} = 10.0 units/Å, and the side chain centroid distance between two disulfide bridge forming cysteines, d_s = 2.71 Å. For crambin simulation, the total energy function is

$$E_{\text{total}} = E_{\text{local}} + E_{\text{nonlocal}} + E_{\text{rg}} + E_{\text{ss}} \qquad (9)$$

The size of the conformational population was set to 200, so that there were 400 conformations in the mutation and crossover population, respectively, and a total of 1000 conformations that were simultaneously computed in each generation of the minimization process.

TABLE 3 Simulations for hybrid apamin sequence, a polypeptide of 18 residues with two disulfide bonds

Parameters	$E_{\rm start}$	$E_{\rm end}$	$E_{\rm C\alpha}$	$E_{\rm SC}$	$E_{\rm S}^{\rm p}$	DME	RMS	Total contacts	Radius of gyration
Average	5949.76	−31.27	72.43	88.69	−166.42	1.70	2.17	234.9	6.21
σ	3911.61	4.95	2.17	4.76	1.62	0.10	0.16	3.8	0.05

200 structures have been computed simultaneously starting from random conformations. The simulation conditions are the same as that in Table 2.

The initial energy profiles of the 200 computed structures have a mean value of 31,850.99 units with a standard deviation of 22,162.93 units (Table 5). The energy profiles of the final optimized 200 structures have a average value of 2084.60 units and a standard deviation of 1.12 units. The lowest energy is 2081.47 units, and the highest is 2085.89 units. The convergence of the energy profile in the genetic algorithm optimization is thus very good. The average DME of the 200 computed structures is 2.29 Å (with the smallest 2.24 Å and the largest 2.34 Å), the average RMS of the computed structures is 2.93 Å (with the smallest 2.82 Å and the largest 3.09 Å). The computed structures have a high degree of structural convergence. The RMS difference between any two of the computed structures is <0.6 Å. All the computed structures have the same topology as that in the crystal structure. The average total contacts of the 200 computed structures are 780 and the average radius of gyration is 9.7 Å. The corresponding values of the crambin crystal structure are 876 and 9.7 Å. Obviously, the computed structures are more compact than that of the crystal structure.

The crystal structure of crambin has two α-helical regions (fragments 7–19 and 23–29) and an antiparallel β-sheet (fragments 1–4 to 32–35), these secondary structure regions in the computed structures appear naturally as a result of structural the optimization. Fig. 6 shows stereo backbone and side chain centroid plot of the crambin crystal structure (Fig. 6 *a*) and one of the computed structures (Fig. 6 *b*). They have very similar folding topology and the secondary structure regions. However, there are qualititive differences in the secondary structure formation between the native and the computed structures. In the computed structures, whereas the second α-helix (fragment 23–29) formed in agreement with the crystal structure, the first α-helix did not form completely; only the fragment residue 6–15 formed correctly, which is about four residues short in comparison with that in the crystal structures. In the β-sheet regions, although the individual fragments (1–4, 32–35) formed a correct β-strand, the distance between these two strands is too far to form proper hydrogen bonding in the computed structures (Fig. 6 *b*). These problems are fixed by the inclusion of the explicit hydrogen bonding energy term in the energy function.

Table 6 summarizes the results for simulation with the explicit hydrogen bonding energy term. The average DME of the computed structure is improved. The average total

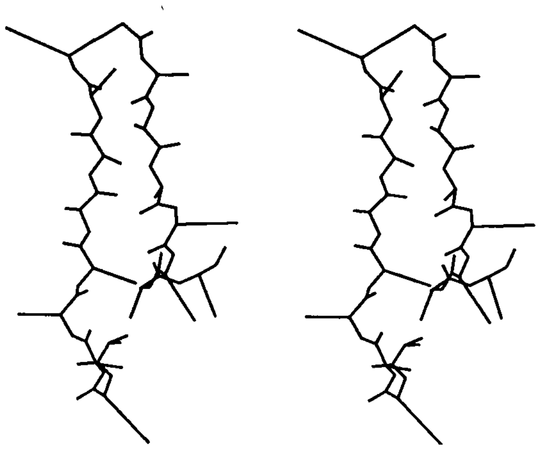

FIGURE 4 Backbone and side chain centroid stereo plot for tachyplesin-I: one of the computed structures.

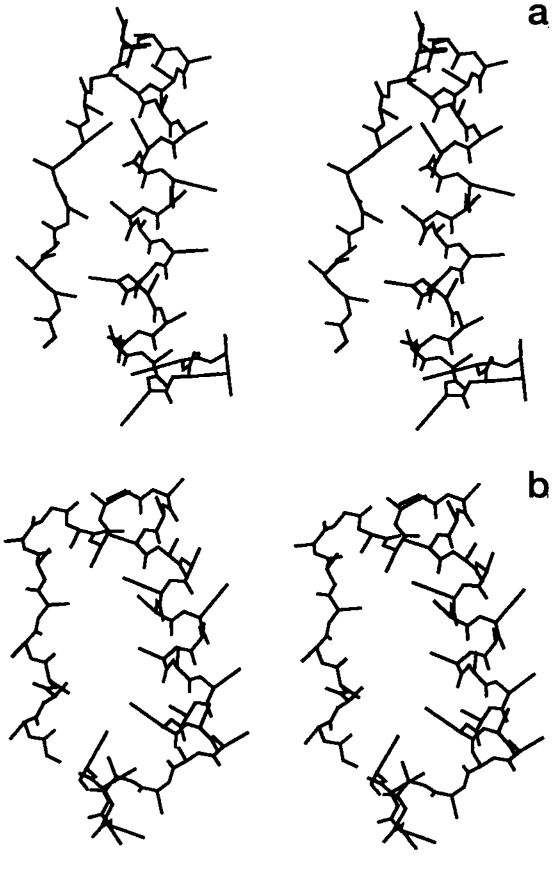

FIGURE 5 Backbone and side chain centroid stereo plot for APPI. (*a*) Crystal structure. (*b*) One of the 90 computed structures (*RMS* = 5.17 Å, *DME* = 3.05 Å to the crystal structure).

TABLE 4 Simulations for APPT, a peptide of 36 residues

Parameters	$E_{\rm start}$	$E_{\rm end}$	$E_{\rm C\alpha}$	$E_{\rm SC}$	$E^{\rm p}_{\rm S}$	DME (Å)	RMS (Å)	Total contacts	Radius of gyration (Å)
Average	4830.82	609.76	671.30	695.17	−687.71	3.05	5.17	448.0	10.5
σ	4083.87	1.21	3.02	2.28	0.29	0.08	0.09	11.2	0.0
Crystal Structure		875.35	786.97	833.94	−661.56			530.0	10.7

90 structures have been computed simultaneously. DME and RMS are computed by using the crystal structure as the reference. σ denotes the standard deviation. The penalty coefficient λ in the $E_{\rm rg}$ has been set to 200 kcal/Å. Hydrogen bond term was used ($E_{\rm hh}$ = −7.0 kcal). The starting conformations were randomly created.

contacts for the computed structure is 833.6, which is much closer to the corresponding value of the crystal structure than that in the previous simulation without the explicit hydrogen bonding energy term. More importantly, all the secondary structure elements are formed correctly in complete agreement with those in the crambin crystal structure. Fig. 6 *c* shows one of the computed structures of this simulation. The RMS difference between the computed structures and the crystal structure is primarily due to the difference in the relative position of C-terminal segment to the β-sheet region, as can be clearly seen from Fig. 6, *a* and *c*. Fig. 6 *d* is the stereo superposition of 30 computed crambin structures and they show a very high degree of structural convergence.

In the computed structures of crambin, all the secondary structure elements (α-helix) formed correctly. The two helical regions folded correctly even without the disulfide bond term, and to a large extent the segment between residues 5 and 30 folded to a native-like conformation (*RMS* < 3.5 Å). The disulfide bond energy term was needed to bring Cys-4 and Cys-32 nearby to form a disulfide bond; otherwise the anti-parallel β-sheet region (fragment 1–4 to 32–35) would not fold properly.

Helix-forming peptides

Several small helices have been designed (Degrado, 1988). We apply our computational method, using as the only input the primary sequence. We find that for these short peptides, including the explicit hydrogen bonding energy term makes little difference to the quality of the results. The (ϕ,ψ) angles of the starting conformations are randomly assigned.

Baldwin peptides

Several peptides have been designed by the Baldwin group (Scholtz and Baldwin, 1992; Marqusee and Baldwin, 1987). We study two 17-mer alanine-based peptides that are known to be highly α-helical in water (Scholtz and Baldwin, 1992; Marqusee and Baldwin, 1987), but for which no specific three-dimensional structures are available. Their sequences are AEAAAKEAAAKEAAAKA and AKAAAEKAAAE-KAAAEA (Marqusee and Baldwin, 1987). The second sequence is just a reversal of the first one. These two sequences have hydrophobicity periodicity, which fits the α-helix conformation. Fig. 7 shows that both Baldwin helix-1 (*top*) and Baldwin helix-2 (*bottom*), are helical. A bend occurs near the N-terminal of the Baldwin helix-2.

α-1

Hill et al. (1990) designed a helix-forming peptide, α-1, and have crystal structures with 2.7 Å resolution. The sequence is GLU-Leu-Leu-Lys-Lys-Leu-Leu-Glu-Glu-Leu-Lys-Gly. Fig. 8 shows the crystal structure of α-1 (*top*) and one of the computed structures (*bottom*). The RMS deviation between these two structures is 1.78 Å. The main difference between the crystal structure and our computed structures is at C-terminal residues 10 and 11, which adopt more extended conformation in the crystal structure than in the computed structures.

Helical erythrocyte lysing peptide (HELP)

HELP is a 26-mer peptide found by NMR experiments (Klaus and Moser, 1992) and CD (Moser, 1992) to be a stable α-helix. Its sequence is GLGTLLTLLEFLLEELLE-FLKRKRQQ. Fig. 9 shows one of the computed structures, which is similar to all the others (RMS <0.5 Å) and similar to the NMR structure of Klaus and Moser (1992; Moser, 1992). Since we do not have the NMR coordinates, detailed comparison with the NMR structure is not possible. Consistent with the NMR results, we find that the N-terminal 3 residues are not α-helical.

TABLE 5 Simulations for crambin, a protein of 46 residues with three disulfide bonds

Parameters	$E_{\rm start}$	$E_{\rm end}$	$E_{\rm C\alpha}$	$E_{\rm SC}$	$E^{\rm p}_{\rm S}$	DME	RMS	Total contacts	Radius of gyration
Average	31850.99	2084.60	1116.43	1322.19	−448.15	2.29	2.93	781.1	9.94
σ	22162.93	1.12	1.79	1.61	0.93	0.03	0.08	1.6	0.04
Crystal Structure		2123.09	1277.21	1290.38	−444.77			876	9.7

200 structures have been computed simultaneously. DME and RMS (unit in Å) are computed by using the crystal structure as the reference. σ denotes the standard deviation. The penalty coefficient λ in the $E_{\rm rg}$ has been set to 200 kcal/Å. An explicit disulfide bridge interaction energy is used (Eq. 8).

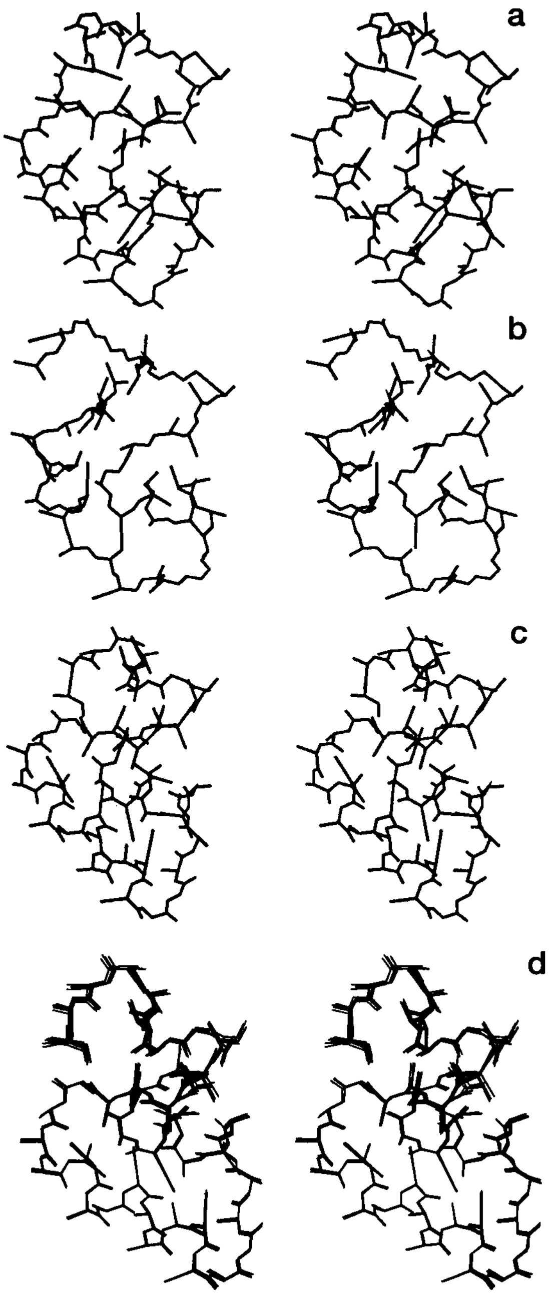

FIGURE 6 Backbone and side chain centroid stereo plot for crambin. (*a*) Crystal structure. (*b*) One of the 200 computed structures (*RMS* = 2.85 Å, *DME* = 2.26 Å to the crystal structure). (*c*) One of the 200 computed structures in minimization with the explicit hydrogen-bonding energy term (*RMS* = 3.03 Å, *DME* = 1.94 Å to the crystal structure). (*d*) Superposition of 30 computed crambin structures; these are all very similar and have an RMS <0.60 Å between any two structures.

Using the primary sequence as the input, we also computed the conformations of the C-peptide, the 13-residue N-terminus fragment of ribonuclease A, which was shown to adopt the α-helical conformation in aqueous solution (Brown and Klee, 1971), and found that it formed α-helix after minimization from random starting conformations (structures are not shown here).

N-terminal of Barnase

Barnase (Sancho et al., 1992, Fersht et al., 1992) has an N-terminal fragment (residues 1–36), which encompasses two α-helices (residues 6–18 and 26–34) in the native structure and forms native-like secondary structure in isolation, as determined by CD and NMR experiments (Sancho et al., 1992). We computed helical probabilities for the 36-residue N-terminal fragment.

The input to the simulation was only the amino acid sequence. Explicit hydrogen bonding was included in the potential function. We performed 10 independent runs, with 200 conformations per run. The α-helix formation probability for a residue was computed as the fractional population in α-helix conformation in all the computed structures normalized by the number of total optimized conformations.

Fig. 10 *a* plots the computed α-helix formation probability for the Barnase N-terminal fragment. Fig. 10 *b* shows experimental results of Sancho et al. (1992), the crystal structure, and a hypothetical nucleation site calculation result by Moult and Unger (1991). Our computed α-helix formation probability agrees in general with both the NMR data of Sancho et al. (1992) and the Barnase crystal structure data; it also covers the region of the hypothetical nucleation site computed by an algorithm that minimizes the surface exposure of hydrophobic residues. Our result indicates that there are two helical segments in conformations that this N-terminal fragment may adopt, α-helix$_1$ (residues 10–23) and α-helix$_2$ (residues 26–33). It is worth pointing out that the α-helix$_2$ forms with a high probability from our calculation, whereas the NMR measurement predicts a much weaker helical formation, and the Moult and Unger (1991) algorithm predicts no helix formation. Our result apparently agrees with the crystal structure data (helix residues 26–34) better.

Zinc finger motif, ubiquitin, and cytochrome 256B

Zinc finger motif, binding a zinc ion through the conserved a pair of Cys and a pair of His residues, is believed to responsible for the DNA-binding activity of the transcription factor (Evans and Hollenberg, 1988). Models of the three-dimensional structure were proposed by homology modeling (Gibson et al., 1988). The NMR-determined structure (Pavletich and Pabo, 1993; Omichinski et al., 1992) indicates that it has an α-helical segment in the C-terminal end and a β-sheet segment in the N-terminal end (Fig. 11 *a*). In the simulation, we only used the primary sequence as the input. The hydrogen-bonding term was also used besides the local and the nonlocal interaction. The computed structures have an average (over 200 computed structures) DME 3.68 Å and an average RMS 5.35 Å. The computed structures have an average radius of gyration of 8.2 Å and a total contact

TABLE 6 Simulations for crambin, a protein of 46 residues with three disulfide bonds

Parameters	E_{start}	E_{end}	$E_{C\alpha}$	E_{SC}	E_S^p	DME	RMS	Total contacts	Radius of gyration
Average	54299.21	1942.24	1148.94	1246.39	−456.52	2.02	3.20	833.6	9.50
σ	39098.91	1.27	2.90	1.45	0.65	0.04	0.07	4.3	0.00
Crystal Structure		2038.09	1277.21	1290.38	−444.77			876	9.7

In addition to all the conditions in the above simulation, the explicit hydrogen bonding energy term is used in this simulation.

number of 418; the corresponding values of the NMR structure are 8.9 Å and 430, respectively. The best computed structure has a DME 3.55 Å and an RMS 5.23 Å compared with the NMR structure. Fig. 11, *b* and *c* shows the stereo plots for two of the computed structures. One can immediately see from the computed structures that the C-terminal α-helix is formed correctly; however, the N-terminal β-sheet between the two strands 2–4 and 10–12 is not formed properly, although these two segments do form a strand-like structure. It seems that there was not enough attractive interaction to bring these two strands together to form a proper β-sheet. This seems a typical phenomenon in our present computational model that was also found in the ubiquitin simulation.

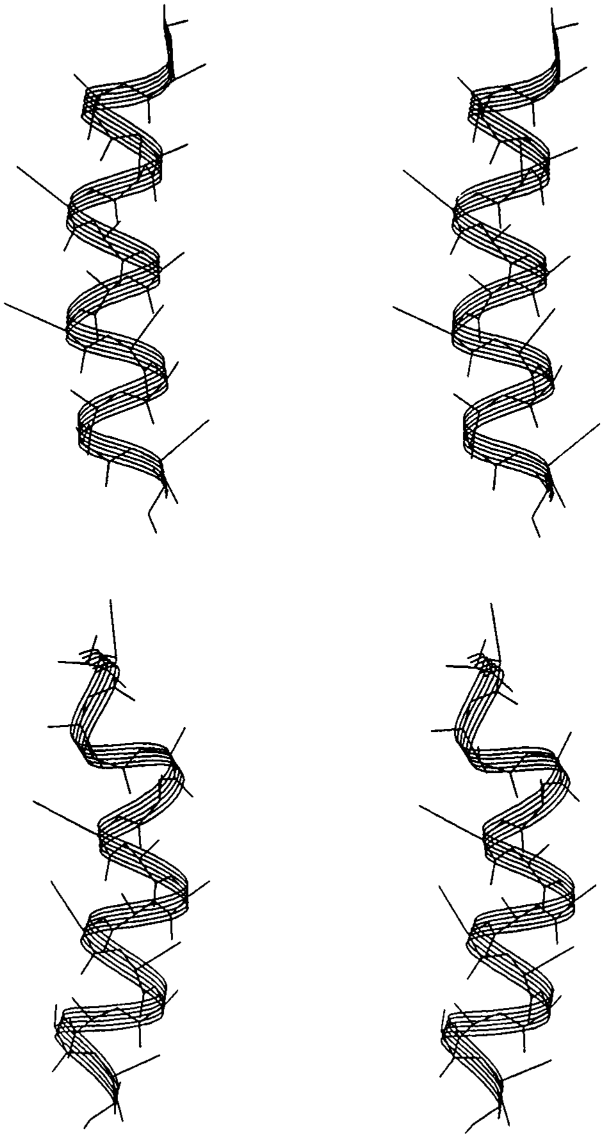

FIGURE 7 Computed structures for Baldwin helices, (*top*) helix-1, (*bottom*) helix-2.

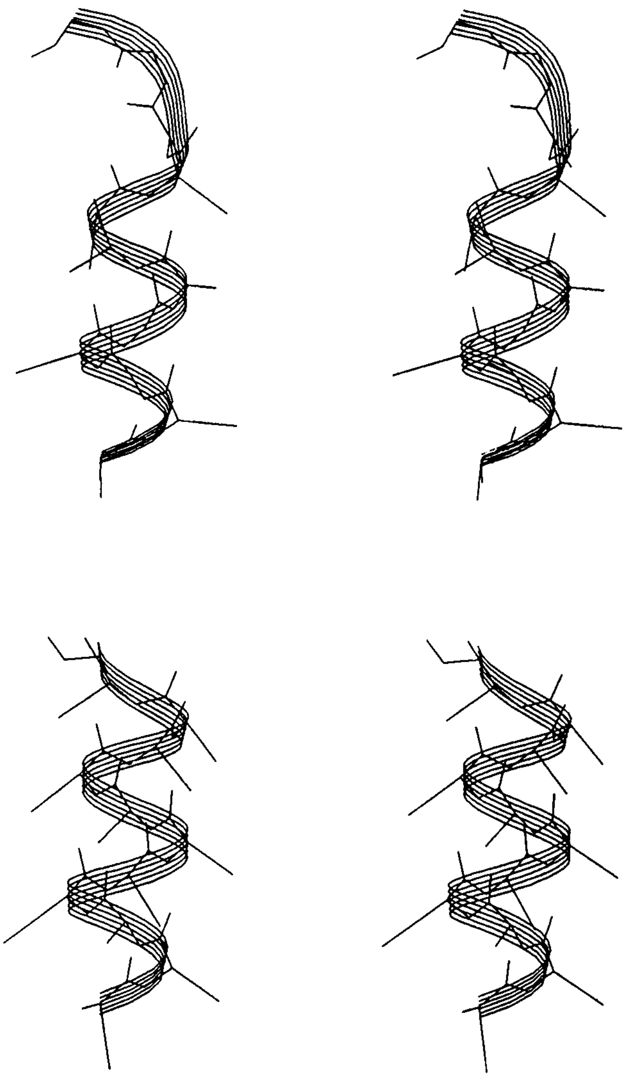

FIGURE 8 Backbone and side chain centroid stereo plot for α-1, (*top*) the crystal structure of α-1, (*bottom*) one of the computed structures (RMS = 1.78 Å).

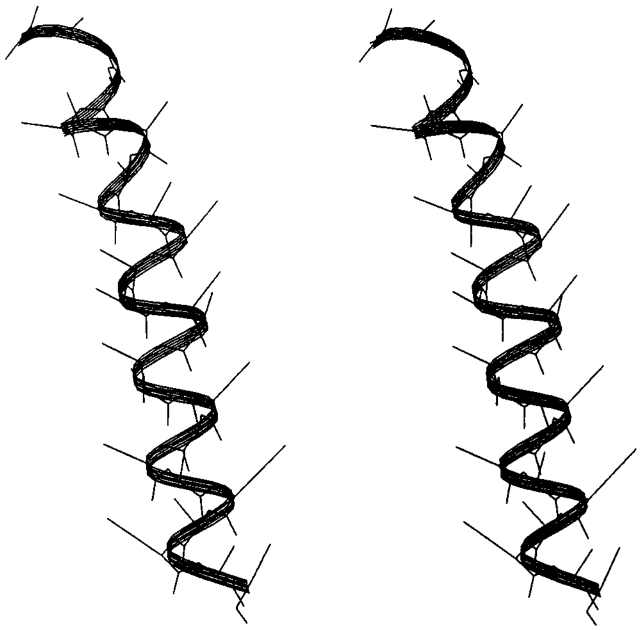

FIGURE 9 One of the computed structures of HELP.

The crystal structure of cytochrome 256b, a 106-residue protein, shows that it is a tightly packed four-helix bundle (Lederer et al., 1991). We used the primary sequence of 256b as the only input in our simulation for cytochrome 256b. The computed structures (Fig. 12) are not as compact as in the crystal structure. This shows that the nonlocal potential is not strong enough to drive the protein to sufficient compactness. Although the folding topology is incorrect in this case, the local secondary structure elements are formed correctly as shown in Fig. 12. When the radius of gyration constraint was introduced, the computed structures were more globular and contained many short α-helices. Because cytochrome 256b is not spherical, the use of an isotropic radius constraint is not helpful in this case.

Native ubiquitin has an α-helix and a β-sheet that consists of five barrel-like β-strands (Vijay-Kumar et al., 1987). Although the computed structures have correct secondary structural elements, the tertiary structures are incorrect. We find that the computed structures have higher final energy than the native structure. This seems to suggest that both the current nonlocal potential function and the current search strategy have their limitations.

CONCLUSIONS

We describe a procedure for computing the structures of peptides and small proteins. It uses a low-resolution chain representation, a genetic algorithm search strategy, and a simple potential function consisting of three parts. 1) The local interactions are modeled using the Biosym Discover force field. 2) The nonlocal interactions are modeled using statistical potentials derived from the Protein Databank. 3) A hydrogen-bonding potential was also tested. In some cases, we also used constraints from disulfides (apamin and crambin) and radius of gyration (melittin, APPI, crambin). The method computes reasonably good structures for several peptides and small proteins covering different architectures, starting from random conformations. The computer time is not excessive (for melittin, the population converges after 42 generations, which takes about 6.0 min on a VAX-6400, and for crambin, the population converges in 58 generations, which takes about 17.0 min) and the genetic algorithm based minimization scheme converses fast (Fig. 13). We do not view our method at its current stage an ab initio method of protein folding. Rather, the simulations described in our present work as well as in our previous work (Sun, 1995; Sun, 1993) focus on the possibility that we can compute a reasonable tertiary structure given the primary sequence of a protein and certain low order constraints. These low order constraints, such as radius of gyration and the S-S bridges, can be obtained from many experiments other than x-ray crystallography. Methods such as these will be useful to understand computationally how protein folds.

APPENDIX

Here is a description of the genetic algorithm. Using the reduced geometric representation adopted for a protein described in section II, a population of conformations (Sun 1993) for a given primary sequence is represented by

$$\left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j}, i \in [1, N]; j \in [1, p] \qquad (10)$$

where $\hbar_n$ is an index for the possible (ϕ,ψ) pairs in the internal coordinate space for a given residue. The index $\hbar_n \in ([-180,+180],[-180,+180])$ for (ϕ,ψ) is a continuous angular variable for a given type of amino acid residue; however it can be made a grid integer variable in the Ramachandran map, or a discrete state variable in the Ramachandran map in which its density distribution is in accordance with the (ϕ,ψ) distribution of the known protein structures. i is the index for residue position along the primary sequence of a protein, while j is the index for the j-th conformation in the conformational population of size p. Different conformations in the population have different local (ϕ,ψ) values for a given primary sequence and each conformation of the protein is characterized by a string structure of $\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)$ and this string structure is considered to be a genetic species in the conformation population which contains a set of $\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)$ string structures as indicated above. The genetic operations in this conformation population are defined as

$$\text{Replication}: \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j} \Rightarrow \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j} \hbar_n \in [\hbar_1, \hbar_T],$$

$$i \in [1, N], j \in [1, p]$$

$$\text{Mutation}: \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j} \Rightarrow \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{j} \textit{at site i.}\ \hbar_n \in [\hbar_1, \hbar_T],$$

$$i \in [1, N], j \in [1, p]$$

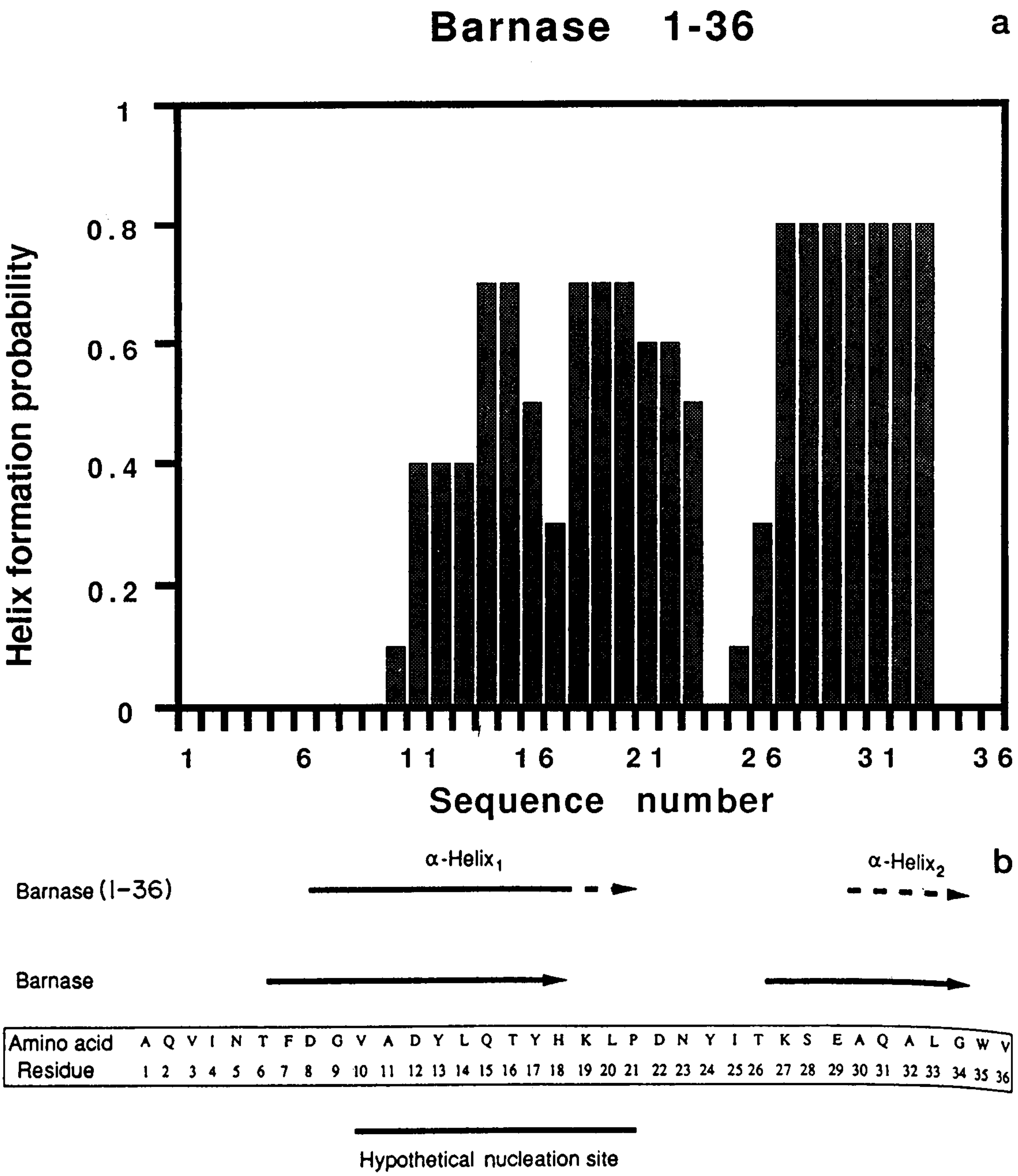

FIGURE 10 Helix formation in the N-terminal fragment (1–36) of Barnase. (*a*) Computed α-helix formation probability. (*b*) Secondary structure of the N-terminal fragment (1–36) in NMR experiment, crystal structure, and the hypothetical nucleation site (adopted from Sancho et al., 1992).

$$\text{Crossover}: \left(\left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j}, \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{l}^{\hbar_m}\right\}^{k}\right) \Rightarrow \left(\left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{j}, \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{l}^{\hbar_n}\right\}^{k}\right)$$

$$\hbar_n \in [\hbar_1, \hbar_T], i, l \in [1, N], j, k \in [1, p] \quad (11)$$

$$\text{with} \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j} \Rightarrow \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_n}\right\}^{k}, \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{j} \Leftarrow \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_m}\right\}^{k},$$

$$i \in [1, \alpha]$$

$$\text{and} \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_n}\right\}^{j} \Rightarrow \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_n}\right\}^{j}, \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{k} \Leftarrow \left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_{i}^{\hbar_m}\right\}^{k},$$

$$i \in [\alpha, N]$$

where $\left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{j}$ denotes the conformations after the mutation operation, $\left(\left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{i}^{\hbar_m}\right\}^{j}, \left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_{l}^{\hbar_n}\right\}^{k}\right)$ denotes the conformations after the crossover operation, the symbols $\Rightarrow$ and $\Leftarrow$ mean copying the corresponding (ϕ,ψ) values in the sites of the target conformations. α is the crossover site for two conformations.

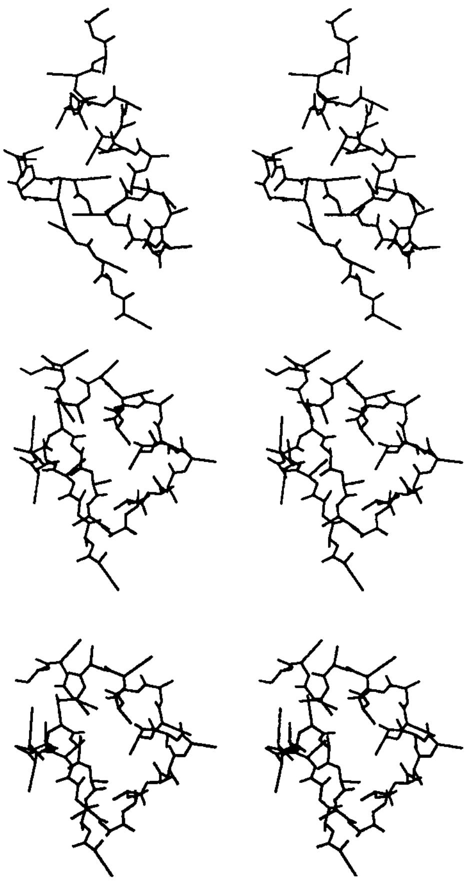

FIGURE 11 Backbone and side chain centroid stereo plot for zinc finger motif. (*a*) The NMR structure. (*b*) One of the 200 computed structures (*RMS* = 5.23 Å, *DME* = 3.55 Å to the NMR structure). (*c*) Another computed structure.

After the three basic genetic operations, a new conformation population is selected from the populations of replication, mutation, and crossover according to their energy profiles:

$$E\left(\left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_i^{\hbar_n}\right\}^j_{\text{replication}},\left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_i^{\hbar_m}\right\}^j_{\text{mutation}},\left\{\begin{pmatrix}\phi'\\ \psi'\end{pmatrix}_i^{\hbar_m}\right\}^j_{\text{crossover}}\right)\Rightarrow\left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_i^{\hbar_n}\right\}^j_{\text{new}}$$

$$\hbar_n \in [\hbar_1, \hbar_T],\ i \in [1, N],\ j \in [1, p] \qquad (12)$$

where E is the reduced potential function.

We have further used the dictionary-assisted segmental mutation conformational search method and the random perturbation procedure (Sun 1993) so that the accessible conformations for a given protein sequence are much greater than the possible combinations of those in the segmental conformations dictionaries: This perturbation procedure can be expressed as

$$A_1^{\alpha}\ldots A_k^{\sigma} :: \begin{pmatrix}\phi + d \cdot z_1\\ \psi + d \cdot z_2\end{pmatrix}_1, \ldots, \begin{pmatrix}\phi + d \cdot z_{2k-1}\\ \psi + d \cdot z_{2k}\end{pmatrix}_k, k = 2, 3, 4, 5 \qquad (13)$$

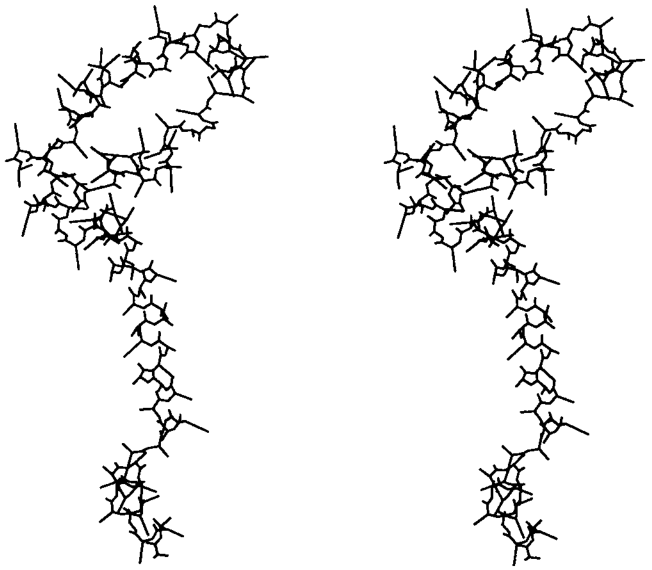

FIGURE 12 One of the computed structures of cytochrome 256b. The simulation input is the primary sequence only. It is not compact; however, the local secondary structure elements seem to agree with the crystal structure, which is a four-helix bundle.

where $z_i \in [-1, +1]$ are uniform random numbers, and d is the range of the perturbation, which is set to be 10° in this study. The conformations generated by the perturbed segmental method satisfy the Ramachandran distribution for each individual amino acid, furthermore these randomly generated conformations have high probability to avoid the local van der Waals conflicts, and therefore effectively reduces the phase space to be searched. The segmentation is carried out randomly, and its probability P_m, m = 2, 3, 4, 5 (di-, tri-, tetra-, pentapeptide segments) satisfies:

$$P_2 + P_3 + P_4 + P_5 = 1.0 \qquad (14)$$

and

$$P_2 > P_3 > P_4 > P_5 > 0.0 \qquad (15)$$

The overall segmentation probability used in the simulations is (P_2, P_3, P_4, P_5) = (0.4, 0.3, 0.2, 0.1), except when noted otherwise. This choice is based on the fact that the larger the probability for the shorter segments, the higher the variability in the constructed conformations. We have used the same segmentation probabilities for the mutation operation in which both of the segmental lengths and the mutation sites in a conformational species are randomly chosen. We have chosen 1 as the number of simultaneous mutation sites for the mutation operation in a conformational species. The partition of the segmentation for any conformational species in different generations is uncorrelated; in other words, we have to repeat the random segmentation for all conformational species in every generation. Termination of the minimization process occurs when no lower energy conformation can be found in 20 consecutive generations of the conformation search.

A share mechanism described in (Sun, 1993) has also been used to forbid premature convergence during the optimization so that a larger region in the conformational space can be searched. We have set the size of the mutation population to be twice as large as the initial population, so the crossover population.

Pittsburgh Supercomputer Center is acknowledged for providing the CPU time (grant DMB93005P). This work is partly supported by Department of Biophysical Science of State University of New York at

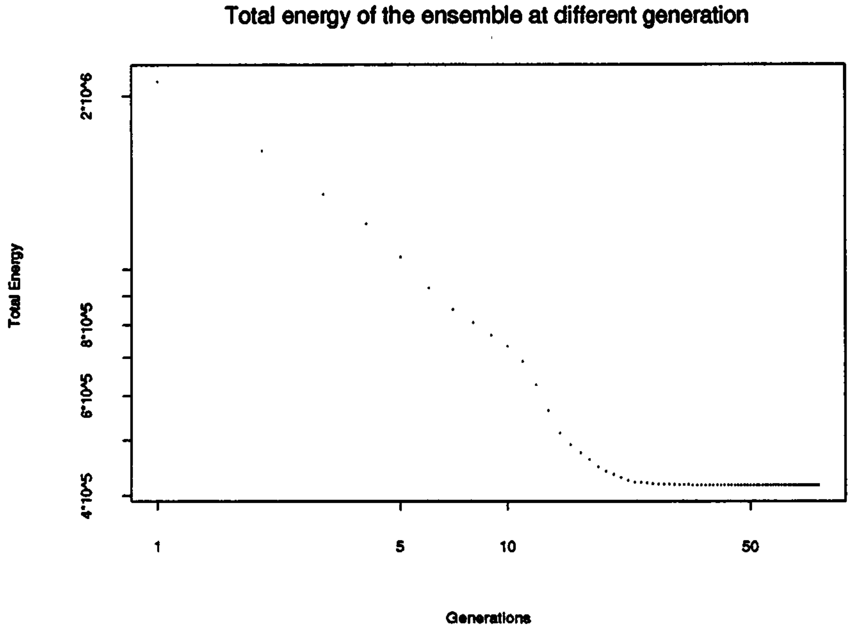

FIGURE 13 Total ensemble energy of the selected new conformation population at each generation during the genetic optimization process (in the case of the crambin simulation). The initial conformation population has a very high energy, which is due to the fact that 1) all the initial conformations are created randomly, and 2) these randomly created structures have high energy because of radius of gyration constraints and the steric hindrance. The total ensemble energy converges to 416,920.7 at the 58th generation and remains unchanged for 20 consecutive generations until the minimization is terminated.

Buffalo, and it is also partly supported by National Institutes of Health and Office of Naval Research grants to Dr. Ken Dill. I thank Dr. Dill for discussions, and Dr. Stanley Burt and Dr. John Erickson for encouragement. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Service, nor does mention of trade names, or organization imply endorsement by the U.S. Government.

REFERENCES

Blommers, M. J. J., C. B. Lucasius, G. Kateman, and R. Kaptein. 1954. Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. *Biopolymers.* 32:45.

Blundell, T. L., J. E. Pitts, I. J. Tickle, S. P. Wood, and W. Wu. 1981. X-ray analysis (1.4-angstrom resolution) of avian pancreatic polypeptide. Small globular protein hormone *Proc. Natl. Acad. Sci. USA.* 78:4175.

Bowie J. U., and D. Eisenberg. 1994. An evolutionary approach to folding small α-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci.* 91:4436–4440.

Brown, J. E., and W. Klee. 1971. Helix-coil transition of the isolated amino terminus of ribonuclease. *Biochemistry.* 10:470–476.

Corey, R. B., and L. Pauling. 1953. Fundamental dimensions of polypeptide chains *Proc. R. Soc. B.* 141, 10

Covell, D. G., and R. L. Jernigen. 1990. Conformations of folded proteins in restricted spaces. *Biochemistry.* 29:3287–3294.

Crippen, G. M., and V. N. Viswanadhan. 1984. A potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* 24: 279–296.

Crippen, G. M., and V. N. Viswanadhan. 1985. Sidechain and backbone potential function for conformational analysis of proteins. *Int. J. Peptide Protein Res.* 25:487–509.

Dandekar, T., and P. Argos. 1992. Potential of genetic algorithms in protein folding and protein engineering simulations. *Protein Eng.* 5:637–645.

Dandekar, T., and P. Argos. 1994. Folding the main chain of small proteins with the genetic algorithm. *J. Mol. Biol.* 236:844–861.

Degrado, W. F. 1988. Design of peptides and proteins. *Adv. Protein Chem.* 39:51–124.

Evans, R. M., and S. M. Hollenberg. 1988. Zinc fingers: gilt by association *Cell.* 52:1–3.

Fersht, A. R., A. Matouschek, and L. Serrano. 1992. The folding of an enzyme. Parts I–VI. *J. Mol. Biol.* 224:771–859.

Fiebig, K. M., and K. A. Dill. 1993. Protein core assembly processes. *J. Chem. Phys.* 98:3475–3487.

Fletcher, R. 1972. *Subroutines for minimization by quasi-Newton methods.* Hartwill report AERE, Hartwell, Berkshire, UK.

Freeman, C. M., C. R. A. Catlow, A. M. Hemmings, and R. C. Hider, 1986. The conformation of apamin. *FEBS Lett.* 197:289–296.

Gibson, T. J., J. P. Postma, R. S. Brown and P. Argos. 1988. A model for the tertiary structure of the 28 residue DNA-binding motif ("zinc finger") common to many eukaryotic transcriptional regulatory proteins. *Protein Eng.* 2:209–218.

Glover, I., I. Haneef, J. Pitts, S. Wood, D. Moss, I. Tickle, and T. Blundell. 1983. Conformational flexibility in a small globular hormone. X-ray analysis of avian pancreatic polypeptide at 0.98-angstroms resolution *Biopolymers.* 22:293–304.

Goldberg, D. 1989. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley Publishing Co., Redding, MA.

Hagler, A. T., and B. Honig. 1978. On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. USA.* 75:554–558.

Hendrickson, W. A., and M. M. Teeter. 1981. Structure of the hydrophobic protein crambin determined directly from the anomalous scattering of sulphur *Nature.* 290:107.

Hill, C. P., D. H. Anderson, L. Wesson, W. F. DeGrado, and D. Eisenberg. 1990. Crystal structure of α 1: implications for protein design. *Science.* 249:543–546.

Hinds, DA, and M. Levitt. 1992. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA.* 89:2536–2540.

Holland, J. H. 1975. Adaptation in natural and artificial system. University of Michigan Press, Ann Arbor, MI.

Kawano, K., T. Yoneya, T. Miyata, K. Yoshikawa, F. Tokunaga, Y. Terada, and S. Iwanaga. 1990. Antimicrobial peptide, tachyplesin I, isolated from hemocytes of the horse-shoe crab (Tachypleus tridentatus). NMR determination of the β-sheet structure. *J. Biol. Chem.* 265: 15365–15367.

Kim P. S., A. Bierzynski, and R. L. Baldwin. 1982. A competing salt-bridge suppresses helix formation by the isolated C-peptide carboxylate of ribonuclease A. *J. Mol. Biol.* 162:187–199.

Klaus, W., and R. Moser. 1992. Nuclear magnetic resonance studies and molecular dynamics simulations of the solution conformation of a "designed," α-helical peptide. *Protein Eng.* 5:333–341.

Kolinski, A., and J. Skolnick 1994. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. II. Application to protein A, ROP, and crambin. *Proteins.* 18:338–366.

Kuntz, I. D., G. M. Crippen, P. A. Kollman, and D. Kimelman. 1976. Calculation of protein tertiary structure. *J. Mol. Biol.* 106:983–994.

Lederer, F., A. Glatigny, P. H. Bethge, H. D. Bellamy, and F. S. Mathews. 1991. Improvement of the 2.5 angstrom resolution model of cytochrome b562 by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* 148:427.

Levitt, M. 1976. A simplified representation of protein conformation for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.

Levitt, M. and A. Warshel. 1975. Computer simulation of protein folding. *Nature.* 253:694–698.

Marqusee, S., and R. Baldwin. 1987. Helix stabilization by Glu-...Lys+ salt bridges in short peptides of de novo design. *Proc. Natl. Acad. Sci. USA.* 84:8898–8902.

Moser, R. 1992. Design, synthesis and structure of an amphipathic peptide with pH-inducible haemolytic activity. *Protein Eng.* 5:323–331.

Moult J., and R. Unger. 1991. An analysis of protein folding pathways. *Biochemistry.* 30:3816–3824.

Omichinski J. G., G. M. Clore, M. Robien, K. Sakaguchi, E. Appella, and A. M. Gronenborn. 1992. High-resolution solution structure of the double Cys2His2 zinc finger from the human enhancer binding protein MBP-1. *Biochemistry.* 31:3907–3917.

Pavletich N. P., and C. O. Pabo. 1993. Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science.* 261: 1701–1707.

Pease, J. H. B., R. W. Storrs, and D. E. Wemmer. 1990. Folding and activity of hybrid sequence, disulfide-stabilized peptides. *Proc. Natl. Acad. Sci. USA.* 87:5643–5647.

Sancho, J., J. L. Neira, and A. R. Fersht. 1992. An N-terminal fragment of barnase has residual helical structure similar to that in a refolding intermediate. *J. Mol. Biol.* 224:749–758.

Scholtz J. M., and R. L. Baldwin. 1992. The mechanism of alpha-helix formation by peptides. *Annu. Rev. Biophys. Biomol. Struct.* 21:95–118.

Sippl, M., M. Hendlich, and P. Lackner. 1992. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: development of strategies and construction of models for myoglobin, lysozyme, and thymosin β_4. *Protein Science.* 1:625–640.

Skolnick, J., and A. Kolinski. 1989. Computer simulations of globular protein folding and tertiary structure. *Annu. Rev. Phys. Chem.* 40: 207–235.

Skolnick, J., and A. Kolinski. 1991. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J. Mol. Biol.* 221:499–531.

Stern, P. S., M. Chorev, M. Goodman, and A. T. Hagler. 1983. Computer simulation of the conformational properties of retro-inverso peptides. I. Empirical force field calculations of rigid and flexible geometries of *N*-acetylglycine-*N'*-methylamide, bis(acetamido) methane, and *N,N'*-dimethylmalonamide and their corresponding C α-methylated analogs. *Biopolymers.* 22:1985–1990.

Sun, S. 1993. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Science.* 2:762–785.

Sun, S. 1995. Reduced representation approach to protein tertiary structure prediction: statistical potential and simulated annealing. *J. Theor. Biol.* 172:13–32.

Sun, S., N. Luo, R. Ornstein, and R. Rein. 1992. Protein structure prediction based on statistical potential. *Biophys. J.* 62:104–106.

Tanaka, S., and H. A. Scheraga. 1976. Medium- and long-range interactions parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules.* 9:945–950.

Terwilliger, T. C., and D. Eisenberg. 1982. The structure of melittin. I. Structure determination and partial refinement. *J. Biol. Chem.* 257:6010.

Tuffrey, P., S. Etchebest, S. Hazout, and R. Levery. 1993. A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn.* 8:1267.

Unger, R., and J. Moult. 1993. Genetic algorithms for protein folding simulations. *J. Mol. Biol.* 231:75–81.

Vijay-Kumar, S., C. E. Bugg, K. D. Wilkinson, R. D. Vierstra, P. M. Hatfield, and W. J. Cook. 1987. Comparison of the three-dimensional structures of human, yeast, and oat ubiquitin. *J. Biol. Chem.* 262: 6396–6399.

Wemmer, D., and N. R. Kallenbach. 1983. Structure of apamin in solution: a two-dimensional nuclear magnetic resonance study. *Biochemistry.* 22:1901–1906.

Wilson, C., and S. A. Doniach. 1989. A computer model to dynamically simulate protein folding: studies with crambin. *Proteins.* 6:193–209.

Yue, K., and K. A. Dill. 1995. Forces of tertiary structural organization in globular proteins. *Proc. Natl. Acad. Sci.* 92:146–150.