Protein Science (1993), *2*, 762-785. Cambridge University Press. Printed in the USA.

Reduced representation model of protein structure prediction: Statistical potential and genetic algorithms

SHAOJIAN SUN

Department of Biophysical Science, State University of New York at Buffalo, Buffalo, New York 14214

(RECEIVED August 18, 1992; REVISED MANUSCRIPT RECEIVED January 4, 1993)

Abstract

A reduced representation model, which has been described in previous reports, was used to predict the folded structures of proteins from their primary sequences and random starting conformations. The molecular structure of each protein has been reduced to its backbone atoms (with ideal fixed bond lengths and valence angles) and each side chain approximated by a single virtual united-atom. The coordinate variables were the backbone dihedral angles ϕ and ψ. A statistical potential function, which included local and nonlocal interactions and was computed from known protein structures, was used in the structure minimization. A novel approach, employing the concepts of genetic algorithms, has been developed to simultaneously optimize a population of conformations. With the information of primary sequence and the radius of gyration of the crystal structure only, and starting from randomly generated initial conformations, I have been able to fold melittin, a protein of 26 residues, with high computational convergence. The computed structures have a root mean square error of 1.66 Å (distance matrix error = 0.99 Å) on average to the crystal structure. Similar results for avian pancreatic polypeptide inhibitor, a protein of 36 residues, are obtained. Application of the method to apamin, an 18-residue polypeptide with two disulfide bonds, shows that it folds apamin to native-like conformations with the correct disulfide bonds formed.

Keywords: conformation population; conformation prediction; genetic algorithms; reduced representation; statistical potential

The basic question of whether it is possible to compute the native conformation of a protein according to certain physical principles with the known physicochemical properties of its constituent amino acids has not yet been answered. Enough evidence has been collected that one can conclude that the primary sequence of a protein determines its structure and function. Furthermore, it has been demonstrated experimentally that the thermodynamic hypothesis that the native structure of a protein is at its global minimum (or deep local minima) of the thermodynamic potential (free energy) of the protein (Anfinsen, 1973) is a valid principle that governs the protein conformational search in both the small perturbation from the native structures of proteins and the processes of denaturing and renaturing for small and medium-size proteins.

The computational difficulties that have hindered progress in protein structure prediction and protein folding come from two related aspects of the problem: first, proteins are highly heterogeneous and have many degrees of freedom, and second, the number of minima in the free energy landscape of the system depends exponentially on the total number of degrees of freedom of the system. Molecular mechanics and molecular dynamics (McCammon & Harvey, 1987; Brooks et al., 1988) with full atomic empirical potential functions have been successfully used in protein studies such as crystallographic structure refinement, normal mode analysis, and free energy simulation, in which the conformations of proteins are near their native states. However, these methods are precluded in studies of protein folding and in large-scale conformational searches due to the computational difficulties. In order

Reprint requests to: Shaojian Sun, Department of Pharmaceutical Chemistry, University of California at San Francisco-Laurel Heights Campus, 3333 California Street, Room 102, San Francisco, California 94118-1204.

to circumvent these two major difficulties, an alternative approach has been developed in which the essential points are reducing the degrees of freedom in amino acids by a simplified geometric representation and smoothing the potential hypersurface by an average potential function over the geometric reduction. The basic assumption of this reduced representation model (RRM) is that the overall folded structure of a protein is relatively insensitive to the fine details of atomic interactions, and a mean field interaction potential should be sufficient to account for the overall folding of a protein. The RRM model will help (1) achieve a higher efficiency in the large-scale conformational search of protein so that meaningful folded protein structures can be generated from the information of the primary sequence only and (2) obtain a better understanding of the basic physics of the protein-folding process.

Various reduced representation models of protein structures have been proposed. Levitt and Warshel (1975) and Levitt (1976) are among the earliest who attempted to use a reduced C^α representation model to compute the near-native conformation of a protein with a starting conformation far away from the native one. Their results on bovine pancreatic trypsin inhibitor (BPTI) have been considered promising. Unfortunately, artificial pulling and pushing potentials have been used in their energy minimization in order to reach the near-native conformation of BPTI. Kuntz et al. (1976), Tanaka and Scheraga (1976), and Hagler and Honig (1978) have also demonstrated the advantages of the reduced representation. Crippen and his coworkers (Oobatake & Crippen, 1981; Crippen & Viswanadhan, 1984, 1985) have derived a C^α–C^α statistical interaction potential function by using information concerning 25 known proteins. Using the C^α statistical potential, they have proven that for small proteins, if starting from native conformations, the proteins tend to stay near their native conformations. Wilson and Doniach (1989) used a full backbone geometric representation to fold crambin starting from a random conformation with C^α and side-chain statistical potential function. More recently, Sippl et al. (1992) used a similar statistical potential to compute the local backbone conformations of several proteins. To further reduce the size of the conformational space to be searched, lattice models of the C^α representation have been employed by a number of authors (Taketomi et al., 1975, 1988; Chan & Dill, 1989a,b, 1990; Lau & Dill, 1989, 1990; Skolnick & Kolinski, 1989; Covell & Jernigan, 1990), and the results have been encouraging in many aspects. The current status of the theoretical understanding of protein structure and protein stability has been excellently reviewed by Chan and Dill (1991, 1993).

In this paper, I present a model of protein structure prediction that has been recently described (Sun et al., 1992; Sun & Luo, in prep.). The geometric representation of a protein is similar to that of Wilson and Doniach (1989; also Sun & Luo, in prep.). The reduced interaction potential function used in the present model is the same statistical potential function as that reported previously (Sun et al., 1992; Sun & Luo, in prep.). The reduced interaction potential function includes singlet and doublet local interactions of amino acids as well as the nonlocal interactions of one to four and above among amino acids along the primary sequence of a protein. Because we are dealing with the problem of folding from the primary sequence, gradient-dependent energy minimization methods are not applicable due to the fact that these methods can only search a very limited subset of conformational space. The simulated annealing optimization method, as an improved Metropolis Monte Carlo search technique, has been found to be not so efficient in the conformational search process for the current RRM (Sun & Luo, in prep.). This is apparently due to the following two reasons: (1) The search is a single path search in the phase space of the system for a given starting conformation. (2) The annealing process has to be very slow in order to reach possible optimal conformations.

In the present study, I have employed concepts from the genetic algorithm optimization method, which will be described in detail later in the text. The genetic algorithms are searching algorithms based on the mechanics of natural selection and natural genetic operations. They were developed initially by John Holland (1975) and his colleagues during the 1960s and 1970s, and have received increased interest and attention in various fields that closely relate to global optimization (Goldberg, 1989). The merits of using the genetic algorithms as energy optimizers in conformational search are: (1) the search is a multipath search in the phase space of the system, and therefore it may improve the convergence of the search greatly; (2) the method is intrinsically parallel and can simultaneously optimize a population of conformations; and (3) it seems possible that the conformational search method based on the genetic algorithms can partially overcome the problem of less sampling in a large-scale conformational search because of the multipath simultaneous search in genetic algorithms, therefore the number of effective searches may be substantially increased. Standard genetic algorithms (Goldberg, 1989) with binary digital coding methods have been used recently to assign side-chain rotamer conformations with the known fixed backbone conformation of a protein (Tuffrey et al., 1991) and to analyze conformations of a dinucleotide photodimer (Blommers et al., 1992). However, the difficulty of computing the tertiary structure from the primary sequence of a protein, even in the reduced representation model, still requires a more robust method of conformational search. In this paper, I present a genetic algorithm-based method of conformational search, which does not utilize binary coding of the parameter space, as is usually done in standard applications, in order to achieve a robust conformation search.

Reduced representation:
Geometry and the potential function

The RRM is motivated by the following considerations:

1. RRM reduces the total number of degrees of freedom of the protein so that the conformational space that has to be searched is reduced exponentially.
2. RRM smooths the potential hypersurface of the protein so that the number of the local potential minima of the protein is also reduced exponentially. The overall topographical complexity is then reduced significantly. Both 1 and 2 reduce the overall complexity of the computation.
3. RRM-computed optimal conformations can serve as the starting conformations for optimization of full atomic representation structures by either conventional molecular mechanics conformational search or molecular dynamics conformational search.

Although many atomic-level details are lost due to smoothing in the RRM scheme, this method allows one to effectively search a much larger portion of conformational space.

Despite having many degrees of freedom, most native proteins adopt a single compact conformation around which local fluctuations may occur. Such frozen structures (Bryngelson & Wolynes, 1990; Shakhnovich & Gutin, 1990) result from a balance between the van der Waals, electrostatic, hydrogen-bonding interactions, as well as strong solvent effects. A useful RRM of protein structure should include the following essential elements: (1) geometric features of proteins and known constraints, (2) solvent effects of amino acids as the major part of the reduced potential function, and (3) proper consideration of the heterogeneity of the solvent effect in different amino acids.

The geometric representation (Wilson & Doniach, 1989; Sun et al., 1992; Sun & Luo, in prep.) of each protein in the current RRM has been set up as follows (Fig. 1):

1. All backbone bond lengths and bond angles are kept at their ideal values.
2. All the peptide bond dihedral angles are fixed in the trans ($\omega = 180°$) conformation.
3. A single virtual atom is used to represent each side chain at the center of mass of the heavy atoms in the side chain.

The geometric variables that determine protein conformation in this representation are the backbone dihedral angles ϕ and ψ.

The potential function adopted for the current RRM consists of two parts (Fig. 1): the local part and the nonlocal part. The local part characterizes the interactions of singlet residues in a mean field potential and interactions between amino acid residues that are neighbors along the primary sequence. The local interactions, which contain predominantly the steric constraints such as those manifested in Ramachandran plots (Ramachandran & Sasisekhharan, 1968), are defined as a function of the backbone dihedral angles (ϕ, ψ). The nonlocal part represents the interactions between residues that are separated in the primary sequence by at least two intervening residues, i.e., at least one to four interactions. The nonlocal interaction, which contains the relatively long distance interactions such as the electrostatic interactions and the short distance van der Waals interaction of residues near in space but far in primary sequence and the effects of statistical thermodynamics such as the hydrophobic effect, is defined as a function of the distance between the nonlocal residue pairs. Therefore, the general form of the interaction potential for the RRM can be written as follows:

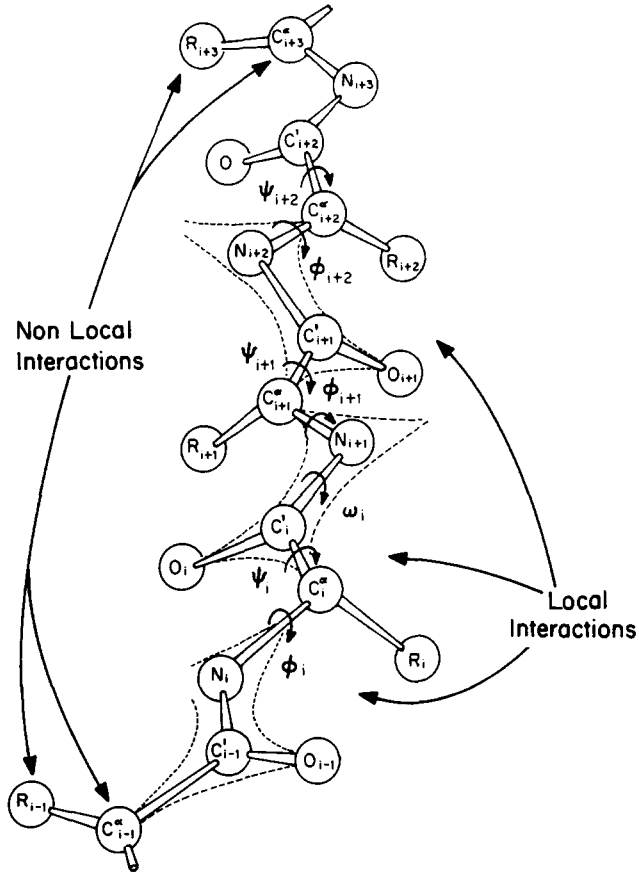

Fig. 1. Current reduced representation model (RRM) of protein structure. Full backbone and side-chain centroids have been preserved in the model; bond lengths and bond angles are held to their ideal values; dihedral angle $\omega = 180°$; the geometric variables are ϕ, ψ at both sides of C^{α}. A side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

$$H = E_{\text{local}}(\{\phi_i, \psi_i\}) + E_{\text{nonlocal}}(\{r_{ij}^{C^\alpha}, r_{ij}^{SC}, r_{ij}^{C^\alpha - SC}\})$$
$$= \alpha_S \sum_i E_{k_i}^S(\phi_i, \psi_i) + \alpha_D \sum_i E_{l_{i-1},k_i}^D(\phi_i, \psi_i) + \ldots$$
$$+ \alpha_{C^\alpha} \sum_{i<j-2} E_{l_i k_j}^{C^\alpha}(r_{ij}^{C^\alpha}) + \alpha_{SC} \sum_{i<j-2} E_{l_i k_j}^{SC}(r_{ij}^{SC})$$
$$+ \alpha_{C^\alpha - SC} \sum_{i<j-2} \left\{ E_{l_i k_j}^{C^\alpha - SC}(r_{ij}^{C^\alpha - SC}) + E_{l_j k_i}^{C^\alpha - SC}(r_{ji}^{C^\alpha - SC}) \right\} + \ldots, \quad (1)$$

where i or j is the position of a residue in the primary sequence, l and k denote the amino acid type, $r_{ij}^{C^\alpha}$ is the distance between the C^α's of the residues i and j, r_{ij}^{SD} is the distance between the centroids of the side chains of the residues i and j, and $r_{ij}^{C^\alpha - SC}$ is the distance between the C^α of residue i and the side-chain centroid of residue j. $E_{k_i}^S$ in Equation 1 is the singlet dihedral angle potential, $E_{l_{i-1}k_i}^D$ the doublet dihedral angle potential, $E_{l_i k_j}^{C^\alpha}$ the potential between the C^α's of residues i and j, $E_{l_i k_j}^{SC}$ the potential between the side chains of i and j, and $E_{l_i k_j}^{C^\alpha - SC}$ the potential between the C^α of i and the side-chain centroid of j. α's are empirical coefficients that represent the relative importance of the various terms in Equation 1.

There are two distinct methods for determining the potential function of the RRM. In the first method, the potential function is obtained by averaging the empirical potential functions of a full atomic representation over the geometric reduction (empirical potential) (Levitt & Warshel, 1975; Levitt, 1976); in the second method, the potential function is derived from information on the known protein structures by statistics (statistical potential) based on the assumption that the statistical distribution of the conformations of the native proteins reflects all possible mean-field interactions and the statistics of geometric constraints (Oobatake & Crippen, 1981; Crippen & Viswanadhan, 1984, 1985; Wilson & Doniach, 1989; Sun et al., 1992; Sun & Luo, in prep.). While the empirical reduced potential function derived by the first method is more acceptable physically, the potential function (statistical potential) derived by the second method is much easier to obtain in terms of the complexity of computation and probably provides a more realistic representation of computationally difficult features such as solvation. I want to emphasize that the statistical potential function derived from the known X-ray structures presumably includes contributions from all kinds of interactions in native structures of proteins. One of the disadvantages of the statistical potential is that one cannot be sure that the model will generate meaningful intermediate folded structures. Because I am here interested only in the final folded conformation, however, the use of a statistical potential is reasonable.

The specific form of the potential function used for this work is as follows:

$$H = \alpha_1 \sum_i E_{k_i}^{(1)}(\phi_i, \psi_i) + \alpha_2 \sum_i E_{l_{i-1},k_i}^{(2)}(\phi_i, \psi_i)$$
$$+ \alpha_3 \sum_{i<j-2} E_{l_i k_j}^{(3)}(r_{ij}^{C^\alpha}) + \alpha_4 \sum_{i<j-2} E_{l_i k_j}^{(4)}(r_{ij}^{SD}). \quad (2)$$

The nonlocal part of the potential in Equation 2 is essentially the same as the potential function used by Wilson and Doniach (1989). This model potential function differs from those used in previous studies by including explicitly the local interaction energy to represent the effects of local steric constraints, which are manifested in the chirality of the backbone structures and the preferred conformations within the regular secondary structures of proteins.

In this study, I have used the statistical method to determine the interaction potential function for the RRM. The nonlocal interaction terms—C^α-C^α and side-chain-side-chain interactions—both include only those residue pairs that are separated and are at least one to four interactions. The density function $\rho(r_{lk})$ for a given pair of lth and kth kinds of amino acids at distance r_{lk} is given by

$$\rho(r_{lk}) \propto \exp[-E_{lk}(r_{lk})/RT], \quad (3)$$

which can be rewritten as

$$E_{lk}(r_{lk}) = -RT \ln\left[\frac{N_{lk}(r_{lk})}{N_0(r_{lk})}\right], \quad (4)$$

where R is the gas constant, T is the temperature, $N_{lk}(r)$ is the distribution of the amino acids of type l at $r = 0$ and of type k in the shell of $r + \delta r$, and $N_0(r)$ is the normalization factor corresponding to the uniform distribution. The value of δr was set to be 1.0 Å. The energy function E_{lk} is calculated statistically from known protein structures selected from the Brookhaven Protein Data Bank (Bernstein et al., 1977; Abola et al., 1978) with a homology score less than 50%. Both the C^α-C^α interaction potential and the side-chain-side-chain potential can be calculated according to Equation 4.

The local interaction terms account for the preferred conformation of singlet and doublet residues. It has been known for a long time that the (ϕ, ψ) distribution for a given amino acid or given pair of amino acids is limited in dihedral angle space (Ramachandran & Sasisekhharan, 1968). The limitation in the (ϕ, ψ) values reflects the local steric hindrance and the local structure preferences in proteins, e.g., those in regular structures. This preference can also be converted into an empirical potential function in (ϕ, ψ) space for each individual amino acid according to the Boltzmann principle. Let $n^k(\phi, \psi)$ denote the sample density function of k type of amino acid in (ϕ, ψ) space according to the known protein structures. Then

$$n^k(\phi, \psi) \propto \exp[-E_k(\phi, \psi)/RT], \quad (5)$$

which may be rewritten as

$$E_k(\phi, \psi) = -RT \ln\left[\frac{n^k_{\delta\phi,\delta\psi}(\phi, \psi)}{N_k \delta\phi \delta\psi}\right], \quad (6)$$

where $\delta\phi$, $\delta\psi$ are the grid sizes in dihedral angle space, k indicates the type of the amino acid, and N_k is the normalization constant, which is equal to the number of total samples in the (ϕ, ψ) space for the amino acid of type k.

Dipeptide conformations $(\phi_{i-1}, \psi_{i-1}; \phi_i, \psi_i)$, found in the Protein Data Bank, can also be used to compute a pairwise potential in dihedral angle space. Because the samples in the known protein structures are very limited, about 24,000 residues in our studies, it is impossible to compute directly a useful potential function in this four-dimensional dihedral angle space. To solve this problem, I have used a conditional distribution of (ϕ_i, ψ_i) in the Ramachandran map. The directionality information of the primary sequence of a protein is carried in the dipeptide segments of that sequence, therefore one can use the probability distribution of the possible conformations for (ϕ_i, ψ_i) of k_i-type amino acid for a given amino acid in front of it (say l_{i-1}-type amino acid). Similar to the singlet potential, the pairwise correlation potential function (doublet potential) is defined as:

$$E_{l_{i-1},k_i}(\phi_i, \psi_i) = -RT \ln\left[\frac{n^{lk}_{\delta\phi_i,\delta\psi_i}(\phi_i, \psi_i)}{N_{lk} \delta\phi_i \delta\psi_i}\right], \quad (7)$$

where $n^{lk}_{\delta\phi_i,\delta\psi_i}(\phi_i, \psi_i)$ is the sample density function in dihedral angle space of the amino acid type k with the amino acid of type l in front of it. N_{lk} is the normalization constant.

In order to reduce the computational demand, a look-up table has been constructed for each term in the potential function of Equation 2. For nonlocal interactions, the last two terms in Equation 2, a cut-off distance of 15 Å has been chosen beyond which the interaction potential energy is set equal to zero. The local interactions of singlet and doublet potentials are calculated in the following way: the Ramachandran (ϕ, ψ) map for a given amino acid is coarsely divided into 60×60 lattices for singlet potential and 36×36 lattices for doublet potential. The singlet potential function, E_k, for an amino acid of type k is given by

$$E_k(l_\phi, l_\psi) = -RT \ln\left[\frac{n^{eff}_{\delta\phi,\delta\psi}(l_\phi, l_\psi)}{N^k \delta\phi \delta\psi}\right] \quad (8)$$

and

$$n^{eff}_{\delta\phi,\delta\psi}(l_\phi, l_\psi) = \sum_{m=0}^{M} \frac{1}{(2m+1)^2} n_{\delta\phi,\delta\psi}(l_\phi \pm m, l_\psi \pm m), \quad (9)$$

where l_ϕ and l_ψ are lattice number indices, the size of the lattice is $(\delta\phi, \delta\psi) = (6°, 6°)$, $RT = 0.6$, N^k is the normalization constant, which is equal to the total number of the samples for the kth type of amino acid. I have introduced a multiple-average method of Equation 9 to smooth the potential surface, which may otherwise be irregular due to the sparse data sample. The second term in the potential function for doublet, $E_{l_{i-1},k_i}(ll_\phi, ll_\psi)$, is computed similarly with the $(\delta\phi, \delta\psi) = (10°, 10°)$. In both singlet and doublet potential calculations, a multiple-average has been used with smoothing order, $M = 2$. Also a truncation density $n^{trun} = 0.02$ has been used where $n^{eff} = 0$. Inclusion of the local interaction terms has proven to be important to improve the predicted protein conformations (Sun & Luo, in prep.). The way in which the potential function is computed leaves one with the freedom to choose a scaling factor for each term in the total potential function (Equation 2). By considering the differences in magnitudes of these terms and their changes when altering protein conformation, I have set $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (3, 3, 1, 1)$ in all of our simulations. This choice of energy scaling factor has been tested for a number of proteins (Sun et al., 1992; Sun & Luo, in prep.) by using the simulated annealing method of conformation search in the current reduced representation model. With this choice of energy scaling the total variation of the local interaction energy is about 20–30% of that of the nonlocal interaction energy. This is also in agreement with what has been found by Dill's group (K.A. Dill, pers. comm.).

The conformational space to be searched, even in the reduced geometric representation, is so huge that any practical computational search can access only a very small subset of the whole phase space. Further lower-order constraints, such as radius of gyration, volume density, accessible surface area, distribution of the hydrophobic and hydrophilic residues, residues that form disulfide bonds, etc. can be used to bias the conformational search. In this study, I have tested the folding effect of the constraint on radius of gyration, which is described as

$$E_{rg} = \lambda(R_g - R_g^{native})^2, \quad (10)$$

and R_g is defined as

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

where λ is a penalty coefficient and $\vec{CM}$ is the coordinates of center of mass. The typical value of λ is set to be around 8 energy units per residue per Ångstrom.

Genetic algorithms and conformation search

Genetic algorithms are searching algorithms (Holland, 1975; Goldberg, 1989) based on the mechanism of natu-

ral selection and natural genetic operations. "They combine survival of the fittest among string structures with a structured yet randomized information exchange to form a search algorithm with some of the innovative flair of human search. In every generation, a new set of string structures is created using bits and pieces of the fittest of the old; an occasional new part is tried for good measure. While randomized, genetic algorithms are not a random walk procedure. They efficiently exploit historical information to speculate on new search points with expected improved performance" (Goldberg, 1989).

This is different from the traditional optimization methods in which a local gradient is used. Standard genetic algorithms differ from normal optimization and search procedures in several aspects:

1. Genetic algorithms require mapping the parameter set into symbolic string structures (either binary digital strings or nonbinary strings).
2. Genetic algorithms simultaneously search a population of points not a single point in the parameter space. The correlation among different points is utilized during optimization process.
3. Genetic algorithms operate the search in the parameter space with probabilistic transition from p points of the current generation to another new p points of the next generation. The transition probabilities of these points are determined by their values of a system objective function.

Quite a few genetic operations have been suggested and computationally tested (Goldberg, 1989), and the most essential genetic operations used in the genetic algorithms are replication, mutation, and crossover. Let us denote a population of representations by a set of string structures $\{P_i^{\hbar_n}\}^j$, $\hbar_n \in [\hbar_1, \hbar_T]$, $i \in [1, N]$, $j \in [1, p]$. $P^{\hbar_n}$ is one of the accessible states at a given position in a string, $\hbar_h$ is the index for a specific state, $\hbar_T$ is the total number of accessible states at a given position in the string structure, i is the index for ith element in the string structure of length N, and j is the index for jth string structure in the total population of p. Let F be the fitness function (or objective function, in our case the energy function) upon which the selection of next generation is based, i.e., $F(\{P_i^{\hbar_n}\}^j)$. The basic genetic operations can be expressed as the following:

Replication: $\{P_i^{\hbar_n}\}^j \Rightarrow \{P_i^{\hbar_n}\}^j$

$$\hbar_n \in [\hbar_1, \hbar_T], i \in [1, N], j \in [1, p].$$

Mutation: $\{P_i^{\hbar_n}\}^j \Rightarrow \{\varphi_i^{\hbar_m}\}^j$ at site i.

$$\hbar_n \in [\hbar_1, \hbar_T], i \in [1, N], j \in [1, p].$$

Crossover: $(\{P_i^{\hbar_n}\}^j, \{P_l^{\hbar_m}\}^k) \Rightarrow (\{\Im_i^{\hbar_m}\}^j, \{\Im_l^{\hbar_n}\}^k$

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

$$\text{with } \{P_i^{\hbar_n}\}^j \Rightarrow \{\Im_i^{\hbar_n}\}^k, \{\Im_i^{\hbar_m}\}^j \Leftarrow \{P_i^{\hbar_m}\}^k \quad i \in [1, \alpha]$$

$$\text{and } \{P_i^{\hbar_n}\}^j \Rightarrow \{\Im_i^{\hbar_n}\}^j, \{\Im_i^{\hbar_m}\}^k \Leftarrow \{P_i^{\hbar_m}\}^k \quad i \in [\alpha, N],$$

where $\{\varphi_i^{\hbar_m}\}^j$ denotes the string structures after the mutation operation, $(\{\Im_i^{\hbar_m}\}^j, \{\Im_l^{\hbar_n}\}^k)$ denotes the string structures after the crossover operation, the symbols $\Rightarrow$ and $\Leftarrow$ mean copying to the corresponding sites into the target string structure, and α is the crossover site for two string structures.

Although I have expressed in Equation 11 one mutation site in one string structure and one crossover site in one pair of string structures, it is possible and sometimes necessary to have several sites in a given string structure mutate simultaneously, and to have several sites in a given pair of string structures cross over simultaneously. In general, the sites of mutation operation in a string structure can be controlled by a probability function ${}^m p_r(\{P_i^{\hbar_n}\}^j)$, which depends on the element position $P_i^{\hbar_n}$ in a string structure. In this study, I used a probability function that is random and uniform in $i \in [1, N]$ in a string structure. Similarly, the site of crossover operation is controlled by a probability function, which is random and uniform in $i \in [1, N]$ in a pair of string structures.

After the three basic genetic operations, a new set of string structures is created which is

$$\{P_i^{\hbar_n}\}^j_{replication}, \{\varphi_i^{\hbar_m}\}^j_{mutation}, \{\Im_i^{\hbar_l}\}^j_{crossover}$$

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

A new population will be selected according to their corresponding values of the fitness function

$$F(\{P_i^{\hbar_n}\}^j_{replication}, \{\varphi_i^{\hbar_m}\}^j_{mutation}, \{\Im_i^{\hbar_l}\}^j_{crossover}) \Rightarrow \{P_i^{\hbar_n}\}^j_{new}$$

$$\hbar_n \in [\hbar_1, \hbar_T], i \in [1, N], j \in [1, p]. \qquad (13)$$

The robustness of the algorithms is demonstrated by a theorem, called the Schema Theorem, proven by Holland (Goldberg, 1989), which states that: Short, low-order, above-average schemata receive exponential increasing trials in subsequent generations.

The genetic algorithms introduced above can be readily used for the purpose of protein conformational search. With the geometric setting introduced in the last section, the following coding scenario will be adopted in the conformation search process in the current RRM. The string structure in a population is now represented by the primary sequence of a protein that one is interested in, and a conformational population is defined such that different conformations in the population have different local (ϕ, ψ) values for a given primary sequence. This conformational population will be expressed as

$$\left\{\begin{pmatrix}\phi\\ \psi\end{pmatrix}_i^{\hbar_n}\right\}^j \quad i \in [1, N], j \in [1, p], \tag{14}$$

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. Then i is the index for residue position along the primary sequence of a protein, whereas j is the index for the jth conformation in the conformational population of size p. In the representation of Equation 14, each conformation of the protein is characterized by a string structure of $\binom{\phi}{\psi}$, and this string structure is considered to be a genetic species in the conformation population, which contains a set of $\binom{\phi}{\psi}$ string structures as indicated above. The simplicity of this protein conformation representation is due to the reduced geometrical representation adopted in the RRM.

The fitness function of the system is now the interaction potential function (or the Hamiltonian of the system) for a protein, which in our case is the statistical potential function defined in the last section.

The three basic genetic operations in the conformational population during the conformational search process are simply defined as:

Replication. The replication operation is simply to copy the whole set of conformations in the last generation of the conformation population to a so-called replication population. Although a probability function of replication can be assigned according to the energy function (fitness function) of a conformation (string structure), in this study I set this probability to unity for all the conformations to be replicated in order to achieve the maximal accessible search.

Mutation. The mutation operation is to change one or simultaneously several local (ϕ, ψ) values in a conformational string structure $\{(\binom{\phi}{\psi}_i^{\hbar_n}\}^j$, $i \in [1, N]$. The mutation operation at a chosen site for a given amino acid residue can be described as the change of $\binom{\phi}{\psi}$ from one point to another point in the Ramachandran map for that specific residue. The nonuniform distributions of (ϕ, ψ) pairs in Ramachandran maps for all 20 amino acids in the known protein structures are a manifestation of the local geometric hindrance and local energetics. It would be wise to utilize these distributions to make the mutation operation more effective. In the practical computation, the (ϕ, ψ) points for an amino acid in the nonhomologous proteins (Table 1) are listed in a look-up table in which each point can be selected uniform randomly. The chosen (ϕ, ψ) is further perturbed randomly by $d \cdot z$, $z \in [-1, +1]$ a uniform random number, and $d = 10°$. $(\phi + d \cdot z_1, \psi + d \cdot z_2)$ is used for the state to which $\{(\binom{\phi}{\psi}_i^{\hbar_n}\}$ is mutated. The mutation conformation population $\{(\binom{\phi}{\psi}_i^{\hbar_n}\}^j$ is generated from the conformation population of the last generation. Each conformation in the population of the preceding generation is locally mutated in the way described above at a given number of sites. Each mutation site $i \in [1, N]$ in a conformation is uniform randomly chosen along the primary sequence. One can generate a mutation population that has a larger size than the replication population. I have set the mutation population to be twice as large as the replication population. I would like to point out that the genetic algorithm conformation search method described here may be more effective than that in which one tries to further code the (ϕ, ψ) parameters by a binary Gray coding, as suggested in the standard application of the genetic algorithms.

Table 1. *List of 110 proteins from which four primary conformational dictionaries have been compiled. They have less than 50% identity in their primary sequences*

1ABP	1BP2	1CC5	1CRN	1CTF	1CTX	1CY3
1ECA	1ETU	1F19	1FDX	1FX1	1GCN	1GCR
1GOX	1HIP	1HMQ	1HNE	1HOE	1LRD	1MBC
1NXB	1PCY	1PFC	1PFK	1PP2	1PPT	1PRC
1RDG	1REI	1RHD	1RNT	1SGT	1SN3	1TEC
1TIM	1TMN	1TNF	1TPP	1UBQ	1UTG	1WSY
2AAT	2ACT	2ALP	2AZA	2CAB	2CCY	2CDV
2CI2	2CNA	2CPP	2CRO	2CYP	2GD1	2GLS
2GN5	2HFL	2INS	2LBP	2LH1	2LIV	2LZ2
2LZM	2MEV	2MLT	2PAB	2PAZ	2PKA	2PLV
2PRK	2RSP	2SGA	2SNS	2SOD	2STV	2TAA
2TMV	2TS1	2WRP	3ADK	3BCL	3C2C	3CLN
3EST	3FAB	3FXC	3GAP	3GRS	3HVP	3ICB
3PGK	3RP2	3SGB	3XIA	451C	4CHA	4CTS
4FD1	4HHB	4RHV	4SBV	5CPA	5CYT	5TNC
6LDH	6LYZ	8ADH	8CAT	9PAP		

Crossover. The crossover operation is to exchange parts of a conformation string structure, Equation 14, according to Equation 11 between a chosen pair of conformations. There are many ways to carry out the crossover operation. In this study, a method of monogamy crossover is adopted. Each conformation can be crossed over only once in a generation, and there is only one crossover site that is uniform randomly chosen along the primary sequence. The crossover conformation population is generated from the replication and mutation conformation populations. The pairs to be crossed over are randomly chosen from the replication and mutation conformation populations.

Assuming the size of the conformation population of last generation to be p, the size of the replication and mutation conformation populations has been set to $2p$, while the size of the crossover conformation population is equal to $2p$, according to the adopted crossover rule. The new

generation of the conformation population of size p is selected from the replication, mutation, and crossover conformation populations of the preceding generation, so that the selected conformations are those that have the lowest energy among the total $5p$ conformational species.

A conformational search algorithm based on the genetic algorithms described above can be formulated as described below.

Conformation search by genetic algorithms

1. Randomly generate the initial conformational population $\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_n}\}_{old}^j$, $i \in [1, N]$, $j \in [1, p]$. Compute the corresponding energy profile $E(\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_n}\}_{old}^j)$;
2. Generate the mutation and crossover conformational populations, $\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_m}\}_{mutation}^j$, $i \in [1, N]$, $j \in [1, p_m]$, and $\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_l}\}_{cross}^j$, $i \in [1, N]$, $j \in [1, p_c]$. Compute the corresponding energy profile $E(\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_m}\}_{mutation}^j$ and $E(\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_l}\}_{cross}^j)$;
3. Select next generation of the conformational population from the conformational species produced by replication, mutation, and crossover according to their energy profiles

$$Min\left[E\left(\left\{\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)_i^{h_n}\right\}_{old}^j\right), E\left(\left\{\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)_i^{h_m}\right\}_{mutation}^j\right), E\left(\left\{\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)_i^{h_l}\right\}_{cross}^j\right)\right] \Rightarrow \left\{\left(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix}\right)_i^{h_k}\right\}_{new}^j$$

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

4. Loop back to 2 until (i) a prescribed number of generations is reached, (ii) there is no further change of the energy profile in the new generation of the conformational population for a prescribed numbers of generations, or (iii) the following condition is satisfied:

$$\frac{\sum E_j^{old} - \sum E_j^{new}}{\sum E_j^{new}} \leq \epsilon. \tag{15}$$

5. Analyze the final conformational population $\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_n}\}_{final}^j$, $i \in [1, N]$, $j \in [1, p]$.

In order to further increase the efficiency of conformational search by the genetic algorithms described above, a conformational dictionary-assisted search method has been developed. This method is based on the assumption that short peptide (dipeptide, tripeptide, tetrapeptide, and pentapeptide) conformations found in known protein structures and small perturbations of these conformations are the most probable conformations in the protein environment. Primary conformation pools have been compiled for di-, tri-, tetra-, and pentapeptides from the known structures of proteins that have less than 50% homology in their primary sequences (Table 1). These proteins are a subset of the protein structures used by Heringa and Argos (1991). These conformation pools are called conformational dictionaries for the particular oligopeptide, and have a format as follows:

Dipeptide	$A_1^\alpha A_2^\beta$	$[(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_1, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_2]$
Tripeptide	$A_1^\alpha A_2^\beta A_3^\gamma$	$[(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_1, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_2, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_3]$
Tetrapeptide	$A_1^\alpha A_2^\beta A_3^\gamma A_4^\delta$	$[(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_1, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_2, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_3, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_4]$
Pentapeptide	$A_1^\alpha A_2^\beta A_3^\gamma A_4^\delta A_5^\epsilon$	$[(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_1, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_2, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_3, (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_4 (\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_5]$

where $A_i^{\alpha,\beta,\gamma,\delta,\epsilon}$ is one of the 20 amino acids. The proteins from which the four dictionaries have been compiled are listed in Table 1. The current sizes of the conformational dictionaries are 23,626 for dipeptides, 23,279 for tripeptides, 22,943 for tetrapeptides, and 22,097 for pentapeptides. In the conformational search process, as an option, a perturbation has been added to each dihedral angle variable so that the conformations accessible to a given protein sequence are much greater than the possible conformational combinations from these four primary dictionaries. This perturbation procedure can be expressed as

$$A_1^\alpha \cdots A_k^\alpha :: \begin{pmatrix}\phi + d\cdot z_1\\ \psi + d\cdot z_2\end{pmatrix}_1, \cdots, \begin{pmatrix}\phi + d\cdot z_{2k-1}\\ \psi + d\cdot z_{2k}\end{pmatrix}_k, \quad k = 2, 3, 4, 5, \tag{16}$$

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. To utilize this dictionary-assisted conformational search by genetic algorithms, a species in the conformation population, $\{(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_n}\}^j$, has to be randomly divided into segmental conformation strings $\{{}^{s_k}(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})_i^{h_n}\}^j$, $i \in [1, N]$, $j \in [1, p]$, and $\sum s_k = N$. The segmental partition for different conformational species in the conformation population is different since the partitioning of the segments is carried out randomly. In this case, the local mutation operation is defined as the replacement of a segmental conformation string structure ${}^{s_k}(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})^{h_n}$ by another segmental conformation string structure ${}^{s_k}(\begin{smallmatrix}\phi\\ \psi\end{smallmatrix})^{h_n}$ of identical primary sequence. The segment s_k has been chosen to be a 2–5 residue peptide segment with a probability P_m, $m = 2, 3, 4, 5$, so that

$$P_2 + P_3 + P_4 + P_5 = 1.0, \tag{17}$$

and

$$P_2 > P_3 > P_4 > P_5 > 0.0. \tag{18}$$

The four conformation dictionaries provide the primary conformation pools from which mutated segmen-

tal conformations are randomly selected. Both the mutation sites and the segmental length of the mutation are randomly chosen. In all four dictionaries, especially those for the di- and tripeptides, there are cases in which the same primary sequence segment corresponds to more than one conformation (several hundred conformations in many cases). The search process will randomly select one such conformation at each step. The use of the dictionary-assisted segmental conformation search prohibits conformations that have local van der Waals conflicts and therefore effectively reduces the phase space to be searched. This method, in combination with the genetic algorithms described above, is very robust, and the convergence rate is high in comparison with the simulated annealing conformation searches, at least in case of the RRM computations. The segmental genetic algorithm conformation search algorithm can be formulated as described below.

Conformation search by segmental genetic algorithms

1. Randomly partition each conformation species into doublet, triplet, quadruplet, and quintuplet segments with probabilities satisfying Equations 17 and 18, and randomly generate the initial conformation population from the primary pools with additional perturbations, $\{{}^{s_k}(\frac{\phi}{\psi})_i^{h_n}\}_{old}^{j}$, $i \in [1, N]$, $j \in [1, p]$; compute the corresponding energy profile $E(\{{}^{s_k}(\frac{\phi}{\psi})_i^{h_n}\}_{old}^{j})$.
2. Generate the mutated and crossed-over species $\{{}^{s_m}(\frac{\phi}{\psi})_i^{h_m}\}_{mutation}^{j}$ and $\{{}^{s_l}(\frac{\phi}{\psi})_i^{h_l}\}_{cross}^{j}$, and compute the corresponding energy profile $E(\{{}^{s_m}(\frac{\phi}{\psi})_i^{h_m}\}_{mutation}^{j})$ and $E(\{{}^{s_l}(\frac{\phi}{\psi})_i^{h_l}\}_{cross}^{j})$. The ${}^{s_m}(\frac{\phi}{\psi})_i^{h_m}$ mutation is randomly chosen from the available segmental conformations in the corresponding dictionary of the identical primary sequence. The crossover site is randomly chosen; after the crossover is done each conformational species will be repartitioned into a new segmental string structure.
3. Select the next generation of conformation population from the conformational species produced by replication, mutation, and crossover operation according to their energy profiles

$$Min\left[E\left(\left\{{}^{s_k}\left(\frac{\phi}{\psi}\right)_i^{h_n}\right\}_{old}^{j}\right), E\left(\left\{{}^{s_m}\left(\frac{\phi}{\psi}\right)_i^{h_m}\right\}_{mutation}^{j}\right), E\left(\left\{{}^{s_m}\left(\frac{\phi}{\psi}\right)_i^{h_l}\right\}_{cross}^{j}\right)\right] \Rightarrow \left\{{}^{s_k}\left(\frac{\phi}{\psi}\right)_i^{h_k}\right\}_{new}^{j}.$$

4. Loop back to 2 until (i) a prescribed number of generations is reached, (ii) there is no further change of the energy profile in the new generation of the conformational population for a prescribed number of generations, or (iii) Equation 15 is satisfied.
5. Analyze the final conformational population $\{{}^{s_k}(\frac{\phi}{\psi})_i^{h_k}\}_{final}^{j}$.

Actually, the segmental genetic algorithm conformational search method is a better use of the genetic algorithms, since one can view the short segmental peptide conformations plus the random perturbation as the highly fitted schemata. By working with them, I have reduced the complexity of the problem by providing the possible partial local solutions of the problem.

It is important to notice, in this algorithm, that I have used the segmental conformations found in known protein structures as the primary conformation pools to randomly create initial conformation population and to replace the local conformational segments by mutation through the genetic operations described above. However, the algorithm can generate local conformations, which may not be found among the known structures of proteins due to the fact that both the mutation sites and the crossover sites are randomly chosen, and the selected primary segmental conformations are perturbed randomly. Of course, the fitted segmental conformations will be preserved during the conformation selection.

In order to avoid the situation in which all conformations in the population converge to a single conformation (this almost always happens!), which may lead to a premature optimization, a share mechanism, which is similar in spirit to that described by Goldberg (1989), has been adopted in the current search algorithm. The share mechanism prohibits the same energy level having more than a prescribed number of conformations (I chose 3). In other words, during the genetic algorithm conformation search process, if many conformations ($n > 3$) correspond to exactly the same energy, only three of these conformations of equal energy will be preserved in the next generation. The selection of the conformations to be retained from this equal energy pool is random. This mechanism slightly decreases the convergence rate (in terms of genetic generations) but enables the system to search a larger region in the conformational space.

In practical computation, the mutation population has been set to be twice as large as the population size of the last generation, so that more new conformations may appear in the conformation population of the next generation provided their energies are among the lowest. The crossover population is set to be the same size as the mutation population. In addition to the random selection of the mutation site, one can select more than one site to simultaneously excise local segmental conformational mutation in a conformational species. I have chosen 1 as the number of sites for simultaneous segmental mutation.

Results

I have applied the RRM genetic algorithm conformation search method to several small proteins; I report here the

computation results on melittin, a protein of 26 residues; avian pancreatic polypeptide inhibitor (APPI), a protein of 36 residues; and apamin, a small protein of 18 residues with two disulfide bonds. The crystal structures for the first two small proteins are known. The crystal structure of melittin (honeybee [*Apis mellifera*] venom) has a resolution of 2.0 Å (Terwilliger & Eisenberg, 1982). The crystal structure of APPI (turkey [*Meleagris gallopavo*] pancreas) has a resolution of 1.37 Å (Blundell et al., 1981; Glover et al., 1983). Attempts to crystallize apamin have so far not been successful, and therefore no X-ray determination of its structure is available. However, NMR spectroscopic data for apamin are available (Wemmer & Kallenbach, 1983). In this study I used apamin backbone dihedral angle data by Freeman et al. (1986) to construct the native apamin structure; this structure was further energy minimized by DISCOVER with BIOSYM force field. I used this structure as the reference to compute the root mean square error (RMS)[1] and distance matrix error (DME) for structures computed by the genetic algorithm conformational search algorithms.

In the computation, the size of the initial conformation population has been set to 90, and the mutation population size is equal to 180. A random monogamy crossover among the species in the conformation population of the last generation and the newly created mutation conformation population is carried out at each generation with a size of 180 conformations in the crossover conformation population. All of the 90 initial conformations were created randomly from the primary segmental conformation pools with additional random perturbation on each variable. The overall segmentation probability is $(P_2, P_3, P_4, P_5) = (0.4, 0.3, 0.2, 0.1)$. 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 the segmental length and the mutation sites in a conformational species are randomly chosen. I have chosen 1 as the number of simultaneous mutation sites for the mutation operation in a conformational species. The partition of the segmentation $\{s_k\}$ for any conformational species in different generations is uncorrelated; in other words, the random segmentation must be repeated for all conformational species in all generations. Termination of the minimization process occurs when no lower energy conformation can be found in 20 consecutive generations of the conformation search. The energy unit is kT in all figures and tables in this paper.

[1] RMS and DME are defined respectively as RMS $= \left\{\frac{1}{N}\sum_i^N (r_i - r_i^c)^2\right\}^{1/2}$, DME $= \left\{\frac{2}{N(N-1)}\sum_{ij}^N (r_{ij} - r_{ij}^c)^2\right\}^{1/2}$, where superscript ^{c} indicates the conformation to which the comparison is made; it is usually the crystallographic conformation of the protein, which in most cases is not far from the native conformation. Distance matrix error, in some cases, serves as a better parameter to compare the overall similarity of two conformations.

Melittin

Table 2 lists the simulation results for melittin. The information about the melittin primary sequence and the radius of gyration of the melittin crystal structure are the only input for the computation. The penalty coefficient λ in Equation 10 was set to 200 energy units/Å. The random perturbation parameter d in Equation 16 was chosen to be 10°. Ninety structures have been optimized simultaneously. The initial total energy profile, E_{start}, of 90 initial structures was very much scattered and ranged from 1,440.08 units to 15,746.34 units. The average total energy of 90 starting structures was 2,912.00 units with statistical standard deviation of 1,960.75 units. On the other hand, the total energy profiles, E_{end}, after the RRM genetic algorithm optimization uniformly converged to their mean value at 1,290.50 units with a standard deviation of 0.31 units. Figure 2 plots the total energy profile (open bar) for all 90 initial conformations and the total energy profile (filled bar) for the conformations after the RRM genetic algorithm energy minimization. The average distance matrix error (DME) and RMS of the 90 computed structures to the crystal structure are 0.99 Å and 1.66 Å. The average total contacts[2] of all 90 optimized structures are 304.9 with a standard deviation of 1.0, and the average value of the radius of gyration is 10.8 Å. The number of total contacts and radius of gyration of all of the 90 computed structures are very close to the corresponding values for the crystal structure, 300 total con-

[2] The contacts were defined for each pair of residues whose C^α were separated by less than 10.0 Å.

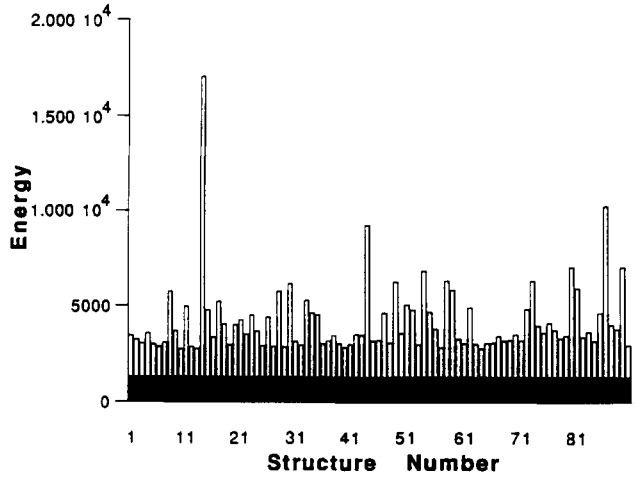

Fig. 2. Energy profile in melittin simulation. Energy profile of the randomly created initial conformations (open bars) and the energy profile for the conformation after the RRM genetic algorithm minimization (filled bars). The average energy of the initial conformations is 2,912.00 units with a standard deviation of 1,960.75 units, and the average energy for the conformation after the minimization is 1,296.50 units with a standard deviation of 0.31 units.

Table 2. *Simulations for melittin, a protein of 26 residues*[a]

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	Total contacts	Radius of gyration
1	2,138.38	1,295.80	308.37	319.94	314.56	352.27	0.99	1.66	306	10.8
2	1,933.50	1,295.80	308.37	319.94	314.56	352.27	0.99	1.66	306	10.8
3	1,761.01	1,295.80	308.37	319.94	314.56	352.27	0.99	1.66	306	10.8
4	2,259.51	1,295.84	309.03	319.66	314.12	352.27	0.99	1.66	304	10.8
5	1,695.33	1,295.84	309.03	319.66	314.12	352.27	0.99	1.66	304	10.8
6	1,564.85	1,295.84	309.03	319.66	314.12	352.27	0.99	1.66	304	10.8
7	1,801.90	1,295.87	308.37	319.94	314.56	352.27	1.00	1.66	306	10.8
8	4,439.09	1,295.87	308.37	319.94	314.56	352.27	1.00	1.66	306	10.8
9	2,379.27	1,295.87	308.37	319.94	314.56	352.27	1.00	1.66	306	10.8
10	1,440.08	1,296.04	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
11	3,667.12	1,296.04	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
12	1,566.87	1,296.04	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
13	1,467.34	1,296.05	309.03	320.15	314.12	352.27	0.98	1.67	306	10.8
14	15,746.34	1,296.05	309.03	320.15	314.12	352.27	0.98	1.67	306	10.8
15	3,472.64	1,296.05	309.03	320.15	314.12	352.27	0.98	1.67	306	10.8
16	2,039.30	1,296.09	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
17	3,926.83	1,296.09	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
18	2,726.00	1,296.09	308.37	320.43	314.56	352.27	0.99	1.67	306	10.8
19	1,654.73	1,296.38	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
20	2,697.42	1,296.38	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
21	2,950.24	1,296.38	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
22	2,211.92	1,296.43	309.03	320.21	314.12	352.27	1.00	1.66	306	10.8
23	3,210.04	1,296.43	309.03	320.21	314.12	352.27	1.00	1.66	306	10.8
24	2,386.07	1,296.43	309.03	320.21	314.12	352.27	1.00	1.66	306	10.8
25	1,622.41	1,296.47	308.37	321.02	314.12	352.27	0.99	1.66	306	10.8
26	3,111.66	1,296.47	308.37	321.02	314.12	352.27	0.99	1.66	306	10.8
27	1,577.53	1,296.47	308.37	321.02	314.12	352.27	0.99	1.66	306	10.8
28	4,462.51	1,296.47	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
29	1,553.65	1,296.47	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
30	4,848.49	1,296.47	308.48	319.67	314.86	352.27	1.02	1.67	306	10.8
31	1,853.37	1,296.50	309.07	319.94	314.56	352.27	0.99	1.65	304	10.8
32	1,663.06	1,296.50	309.07	319.94	314.56	352.27	0.99	1.65	304	10.8
33	4,000.40	1,296.50	309.07	319.94	314.56	352.27	0.99	1.65	304	10.8
34	3,334.50	1,296.54	308.48	320.16	314.86	352.27	1.01	1.68	306	10.8
35	3,226.98	1,296.54	308.48	320.16	314.86	352.27	1.01	1.68	306	10.8
36	1,739.41	1,296.54	308.48	320.16	314.86	352.27	1.01	1.68	306	10.8
37	1,921.52	1,296.54	309.73	319.66	314.12	352.27	0.99	1.65	304	10.8
38	2,159.64	1,296.54	309.73	319.66	314.12	352.27	0.99	1.65	304	10.8
39	1,750.67	1,296.54	309.73	319.66	314.12	352.27	0.99	1.65	304	10.8
40	1,559.57	1,296.57	309.07	319.94	314.56	352.27	1.00	1.65	304	10.8
41	1,720.02	1,296.57	309.07	319.94	314.56	352.27	1.00	1.65	304	10.8
42	2,225.75	1,296.57	309.07	319.94	314.56	352.27	1.00	1.65	304	10.8
43	2,183.84	1,296.58	308.37	321.39	314.12	352.27	0.98	1.68	306	10.8
44	7,944.66	1,296.58	308.37	321.39	314.12	352.27	0.98	1.68	306	10.8
45	1,890.22	1,296.58	308.37	321.39	314.12	352.27	0.98	1.68	306	10.8
46	1,923.63	1,296.59	309.25	319.07	315.41	352.27	0.98	1.64	304	10.8
47	3,355.59	1,296.59	309.25	319.07	315.41	352.27	0.98	1.64	304	10.8
48	1,798.44	1,296.59	309.25	319.07	315.41	352.27	0.98	1.64	304	10.8
49	4,964.73	1,296.59	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8
50	2,273.28	1,296.59	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8
51	3,778.46	1,296.59	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8
52	3,504.84	1,296.63	309.91	318.79	314.97	352.27	0.98	1.64	304	10.8
53	1,690.02	1,296.63	309.91	318.79	314.97	352.27	0.98	1.64	304	10.8
54	5,561.54	1,296.63	309.91	318.79	314.97	352.27	0.98	1.64	304	10.8
55	3,396.25	1,296.63	309.91	318.79	314.97	352.27	0.98	1.65	304	10.8
56	2,522.43	1,296.63	309.91	318.79	314.97	352.27	0.98	1.65	304	10.8
57	1,534.55	1,296.63	309.91	318.79	314.97	352.27	0.98	1.65	304	10.8
58	5,028.48	1,296.66	309.25	319.07	315.41	352.27	0.99	1.64	304	10.8
59	4,539.23	1,296.66	309.25	319.07	315.41	352.27	0.99	1.64	304	10.8
60	2,001.04	1,296.66	309.25	319.07	315.41	352.27	0.99	1.64	304	10.8
61	1,755.71	1,296.66	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8

(*continued*)

Table 2. *Continued*

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	Total contacts	Radius of gyration
62	3,627.62	1,296.66	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8
63	1,724.27	1,296.66	309.25	319.07	315.41	352.27	0.99	1.65	304	10.8
64	1,475.98	1,296.73	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
65	1,777.25	1,296.73	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
66	1,796.56	1,296.73	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
67	2,127.48	1,296.74	308.48	320.29	314.86	352.27	1.01	1.68	306	10.8
68	1,888.86	1,296.74	308.48	320.29	314.86	352.27	1.01	1.68	306	10.8
69	1,922.87	1,296.74	308.48	320.29	314.86	352.27	1.01	1.68	306	10.8
70	2,216.91	1,296.75	309.73	320.15	314.12	352.27	0.98	1.67	304	10.8
71	1,911.39	1,296.75	309.73	320.15	314.12	352.27	0.98	1.67	304	10.8
72	3,552.26	1,296.75	309.73	320.15	314.12	352.27	0.98	1.67	304	10.8
73	5,018.49	1,296.79	308.37	320.90	314.56	352.27	1.00	1.66	304	10.8
74	2,671.00	1,296.79	308.37	320.90	314.56	352.27	1.00	1.66	304	10.8
75	2,310.92	1,296.79	308.37	320.90	314.56	352.27	1.00	1.66	304	10.8
76	2,817.48	1,296.79	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
77	2,449.02	1,296.79	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
78	2,008.17	1,296.79	309.07	320.43	314.56	352.27	0.99	1.67	304	10.8
79	2,139.44	1,296.84	309.25	319.56	315.41	352.27	0.98	1.66	304	10.8
80	5,738.88	1,296.84	309.25	319.56	315.41	352.27	0.98	1.66	304	10.8
81	4,608.19	1,296.84	309.25	319.56	315.41	352.27	0.98	1.66	304	10.8
82	2,064.73	1,296.84	309.25	319.56	315.41	352.27	0.98	1.65	304	10.8
83	2,342.82	1,296.84	309.25	319.56	315.41	352.27	0.98	1.65	304	10.8
84	1,885.61	1,296.84	309.25	319.56	315.41	352.27	0.98	1.65	304	10.8
85	3,344.30	1,296.85	309.03	320.91	314.12	352.27	0.99	1.68	306	10.8
86	8,959.83	1,296.85	309.03	320.91	314.12	352.27	0.99	1.68	306	10.8
87	2,717.67	1,296.85	309.03	320.91	314.12	352.27	0.99	1.68	306	10.8
88	2,489.05	1,296.86	309.91	319.28	314.97	352.27	0.98	1.66	304	10.8
89	5,725.82	1,296.86	309.91	319.28	314.97	352.27	0.98	1.66	304	10.8
90	1,653.66	1,296.86	309.91	319.28	314.97	352.27	0.98	1.66	304	10.8
Average	2,912.00	1,296.50	308.99	319.92	314.69	352.27	0.99	1.66	304.9	10.8
σ	1,960.75	0.31	0.51	0.66	0.46	0.00	0.01	0.01	1.0	0.0

[a] Ninety 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, E_{C^α}, E_{SC}, E_S, E_D are the energy components of C^α-C^α interaction, side-chain–side-chain interaction, singlet interaction and doublet interaction, respectively, of the minimized structures. DME and RMS (units 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 units/Å. The starting conformations were randomly created. The crystal structure has 300 total contacts and a radius of gyration 11.1 Å.

tacts and 11.1 Å for the radius of gyration. The computed structures converge uniformly not only in their final energy but more importantly in their three-dimensional conformations. There is a high degree of similarity among the 90 computed structures. The RMS between any two of the 90 computed structures is less than 0.15 Å for melittin. Figure 3 shows a stereo plot of the backbone and side-chain centroid for the melittin crystal structure and one structure computed by the current model. In all of the computed structures, there is a bending at the middle of the structure due to a proline residue (Pro-14), which is in agreement with the melittin crystal structure. The folding around the N-terminal end in the computed melittin structures is not as helical as that shown in the crystal structure of melittin. The same has been found also in the simulation of melittin described below.

I also tested the current model with the overall segmentation probability of $P_2 = 1.0$ and $P_3 = P_4 = P_5 = 0$ for melittin, that is, only the random doublet segmentation along the chain is used. Under the same conditions given in the above except for the segmentation probability ($P_2 = 1.0$), I obtained similar simulation results, listed in Table 3. Not surprisingly, the initial randomly constructed 90 conformations have very high energy ranging from 1,571.40 to 9,162.78 units. The average value of total energy of 90 initial structures E_{start} is 3,377.8 units with a standard deviation of 1,667.01 units. The 90 optimized structures have an average total energy of 1,264.99 units with a standard deviation of 2.84 units. It is understandable that the sole use of doublet segmentation increases the variability of the conformations that compose the population, in comparison with the case in which the

Table 3. *Simulations for melittin,* $P_2 = 1.0$[a]

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	Total contacts	Radius of gyration
1	1,925.98	1,256.27	304.38	324.29	284.81	342.66	1.69	3.04	276	10.8
2	1,797.16	1,256.27	304.38	324.29	284.81	342.66	1.69	3.04	276	10.8
3	2,104.29	1,256.27	304.38	324.29	284.81	342.66	1.69	3.04	276	10.8
4	2,668.60	1,258.58	304.38	324.29	286.32	343.45	1.69	3.04	276	10.8
5	2,739.85	1,258.58	304.38	324.29	286.32	343.45	1.69	3.04	276	10.8
6	3,473.85	1,258.58	304.38	324.29	286.32	343.45	1.69	3.04	276	10.8
7	2,875.93	1,260.28	309.37	324.83	284.64	341.31	1.74	3.15	272	10.8
8	2,814.33	1,260.28	309.37	324.83	284.64	341.31	1.74	3.15	272	10.8
9	3,575.91	1,260.28	309.37	324.83	284.64	341.31	1.74	3.15	272	10.8
10	2,287.22	1,262.18	305.28	326.11	286.57	343.09	1.75	3.04	276	10.8
11	1,571.40	1,262.18	305.28	326.11	286.57	343.09	1.75	3.04	276	10.8
12	2,225.61	1,262.18	305.28	326.11	286.57	343.09	1.75	3.04	276	10.8
13	1,799.83	1,262.59	309.37	324.83	286.16	342.10	1.74	3.15	272	10.8
14	1,575.42	1,262.59	309.37	324.83	286.16	342.10	1.74	3.15	272	10.8
15	3,874.05	1,262.59	309.37	324.83	286.16	342.10	1.74	3.15	272	10.8
16	1,927.42	1,262.77	299.86	321.64	283.49	350.07	1.78	2.86	282	10.7
17	4,189.52	1,262.77	299.86	321.64	283.49	350.07	1.78	2.86	282	10.7
18	6,255.49	1,262.77	299.86	321.64	283.49	350.07	1.78	2.86	282	10.7
19	4,418.23	1,263.15	304.84	321.90	283.32	348.73	1.78	2.89	278	10.7
20	2,188.64	1,263.15	304.84	321.90	283.32	348.73	1.78	2.89	278	10.7
21	3,288.44	1,263.15	304.84	321.90	283.32	348.73	1.78	2.89	278	10.7
22	1,959.29	1,263.56	303.48	324.29	285.55	350.07	1.59	2.78	282	10.8
23	4,177.65	1,263.56	303.48	324.29	285.55	350.07	1.59	2.78	282	10.8
24	2,921.41	1,263.56	303.48	324.29	285.55	350.07	1.59	2.78	282	10.8
25	4,325.12	1,264.49	305.28	326.11	288.09	343.88	1.75	3.04	276	10.8
26	3,181.17	1,264.49	305.28	326.11	288.09	343.88	1.75	3.04	276	10.8
27	2,519.80	1,264.49	305.28	326.11	288.09	343.88	1.75	3.04	276	10.8
28	2,595.90	1,264.55	305.28	321.36	286.34	350.51	1.49	2.62	278	10.9
29	2,127.36	1,264.55	305.28	321.36	286.34	350.51	1.49	2.62	278	10.9
30	2,548.71	1,264.55	305.28	321.36	286.34	350.51	1.49	2.62	278	10.9
31	6,122.41	1,264.68	301.91	325.03	285.21	350.51	1.73	3.04	288	10.8
32	4,583.71	1,264.68	301.91	325.03	285.21	350.51	1.73	3.04	288	10.8
33	2,663.04	1,264.68	301.91	325.03	285.21	350.51	1.73	3.04	288	10.8
34	2,470.44	1,264.75	304.38	323.72	285.38	351.09	1.68	3.03	276	10.8
35	8,018.83	1,264.75	304.38	323.72	285.38	351.09	1.68	3.03	276	10.8
36	5,811.17	1,264.75	304.38	323.72	285.38	351.09	1.68	3.03	276	10.8
37	5,293.39	1,264.79	305.28	323.04	284.89	351.09	1.69	3.02	276	10.8
38	1,787.76	1,264.79	305.28	323.04	284.89	351.09	1.69	3.02	276	10.8
39	6,024.36	1,264.79	305.28	323.04	284.89	351.09	1.69	3.02	276	10.8
40	5,486.96	1,265.08	299.86	321.64	285.00	350.87	1.78	2.86	282	10.7
41	1,723.92	1,265.08	299.86	321.64	285.00	350.87	1.78	2.86	282	10.7
42	4,206.81	1,265.08	299.86	321.64	285.00	350.87	1.78	2.86	282	10.7
43	3,481.29	1,265.46	304.84	321.90	284.84	349.52	1.78	2.89	278	10.7
44	1,754.54	1,265.46	304.84	321.90	284.84	349.52	1.78	2.89	278	10.7
45	4,037.17	1,265.46	304.84	321.90	284.84	349.52	1.78	2.89	278	10.7
46	3,825.35	1,265.85	304.38	324.29	286.68	350.36	1.69	3.04	276	10.8
47	6,005.03	1,265.85	304.38	324.29	286.68	350.36	1.69	3.04	276	10.8
48	3,160.98	1,265.85	304.38	324.29	286.68	350.36	1.69	3.04	276	10.8
49	1,939.42	1,265.86	303.48	324.29	287.06	350.87	1.59	2.78	282	10.8
50	2,028.87	1,265.86	303.48	324.29	287.06	350.87	1.59	2.78	282	10.8
51	2,626.93	1,265.86	303.48	324.29	287.06	350.87	1.59	2.78	282	10.8
52	1,716.55	1,266.02	310.27	326.65	286.40	341.74	1.79	3.15	272	10.8
53	5,914.45	1,266.02	310.27	326.65	286.40	341.74	1.79	3.15	272	10.8
54	3,175.64	1,266.02	310.27	326.65	286.40	341.74	1.79	3.15	272	10.8
55	1,746.37	1,266.63	308.44	323.36	284.72	349.74	1.73	3.12	274	10.8
56	5,076.25	1,266.63	308.44	323.36	284.72	349.74	1.73	3.12	274	10.8
57	3,127.05	1,266.63	308.44	323.36	284.72	349.74	1.73	3.12	274	10.8
58	3,097.51	1,266.82	305.45	325.04	283.50	349.79	1.81	3.10	274	10.7
59	3,832.12	1,266.82	305.45	325.04	283.50	349.79	1.81	3.10	274	10.7
60	3,525.18	1,266.82	305.45	325.04	283.50	349.79	1.81	3.10	274	10.7
61	2,561.41	1,266.85	305.28	321.36	287.86	351.30	1.49	2.62	278	10.9

(*continued*)

Table 3. *Continued*

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	Total contacts	Radius of gyration
62	3,631.62	1,266.85	305.28	321.36	287.86	351.30	1.49	2.62	278	10.9
63	1,930.24	1,266.85	305.28	321.36	287.86	351.30	1.49	2.62	278	10.9
64	2,866.74	1,266.99	301.91	325.03	286.72	351.30	1.73	3.04	288	10.8
65	3,051.44	1,266.99	301.91	325.03	286.72	351.30	1.73	3.04	288	10.8
66	1,577.12	1,266.99	301.91	325.03	286.72	351.30	1.73	3.04	288	10.8
67	4,451.15	1,267.03	308.47	324.44	285.38	348.73	1.61	2.86	278	10.9
68	1,761.06	1,267.03	308.47	324.44	285.38	348.73	1.61	2.86	278	10.9
69	1,712.63	1,267.03	308.47	324.44	285.38	348.73	1.61	2.86	278	10.9
70	1,676.59	1,267.06	304.38	323.72	286.90	351.88	1.68	3.02	276	10.8
71	2,117.77	1,267.06	304.38	323.72	286.90	351.88	1.68	3.02	276	10.8
72	3,196.91	1,267.06	304.38	323.72	286.90	351.88	1.68	3.02	276	10.8
73	4,947.02	1,267.10	305.28	323.04	286.41	351.88	1.69	3.02	276	10.8
74	2,476.38	1,267.10	305.28	323.04	286.41	351.88	1.69	3.02	276	10.8
75	4,526.70	1,267.10	305.28	323.04	286.41	351.88	1.69	3.02	276	10.8
76	4,088.10	1,267.82	302.77	324.61	286.97	350.95	1.71	2.97	288	10.8
77	2,517.08	1,267.82	302.77	324.61	286.97	350.95	1.71	2.97	288	10.8
78	3,683.91	1,267.82	302.77	324.61	286.97	350.95	1.71	2.97	288	10.8
79	2,052.23	1,267.97	305.16	322.11	283.27	357.20	1.51	2.66	276	10.9
80	2,319.37	1,267.97	305.16	322.11	283.27	357.20	1.51	2.66	276	10.9
81	3,867.83	1,267.97	305.16	322.11	283.27	357.20	1.51	2.66	276	10.9
82	2,299.27	1,267.97	302.77	320.00	286.74	358.37	1.47	2.72	290	10.9
83	8,191.22	1,267.97	302.77	320.00	286.74	358.37	1.47	2.72	290	10.9
84	1,617.04	1,267.97	302.77	320.00	286.74	358.37	1.47	2.72	290	10.9
85	3,188.41	1,268.08	309.37	323.63	285.21	349.74	1.73	3.13	272	10.8
86	9,162.78	1,268.08	309.37	323.63	285.21	349.74	1.73	3.13	272	10.8
87	8,683.59	1,268.08	309.37	323.63	285.21	349.74	1.73	3.13	272	10.8
88	4,488.94	1,268.33	310.27	326.65	287.92	342.54	1.79	3.15	272	10.8
89	2,705.15	1,268.33	310.27	326.65	287.92	342.54	1.79	3.15	272	10.8
90	2,452.05	1,268.33	310.27	326.65	287.92	342.54	1.79	3.15	272	10.8
Average	3,377.77	1,264.99	305.18	323.77	285.75	348.84	1.69	2.96	278.0	10.8
σ	1,667.01	2.84	2.75	1.67	1.33	4.29	0.10	0.16	5.1	0.1

[a] Simulation is carried out under the same conditions as that in Table 2, except that the random segmentation probability is different. In this simulation, only the doublet segmentation has been used in creating the initial conformation population and in the mutation operation. $P_2 = 1.0$, and $P_3 = P_4 = P_5 = 0.0$

segmentation probabilities of longer segments are not equal to zero. In the 90 optimized structures, the structures closest to the melittin crystal structure have a DME of 1.47 Å, and an RMS of 2.72. The average DME for 90 structures is 1.69 Å, and the average RMS is 2.96 Å. It is interesting to note that the optimized structures, when only doublet segmentation is used, are slightly less compact than the simulation results in Table 2. The average total contacts in this case is 278, which is less than the 300 in the crystal structure. Also as expected, the final structures are very similar to each other; however, the optimization took longer time (about one-third more time in comparison to the simulation in Table 1) to converge to the final conformation population.

I would like to point out that the average total energy of the computed structures in this simulation is about 2.5% lower than that in Table 2. This is apparently a typical problem inherent in many computational models of protein structure, especially reduced representation models: the actual native state may not be the lowest energy state (Levitt, 1976; Wilson & Doniach, 1989; Covell & Jernigan, 1990; Sun & Luo, in prep.). A protein system, which contains hundreds or even thousands of atoms, has many degrees of freedom, and its low-lying energy levels correspond to a nearly continuous spectra. Even if the native state of a protein is at the global minimum of its energy landscape, any practically useful computational model of protein structure may deform the shape of the energy landscape of the system due to the approximations introduced in the interaction potential function. Therefore, it is not surprising that the lowest energy conformation may not correspond to the native structure of the protein in many computational models including the full atomic representation model in which an empirical potential function is used. In this case, I believe that this problem originates in the RRM potential function in

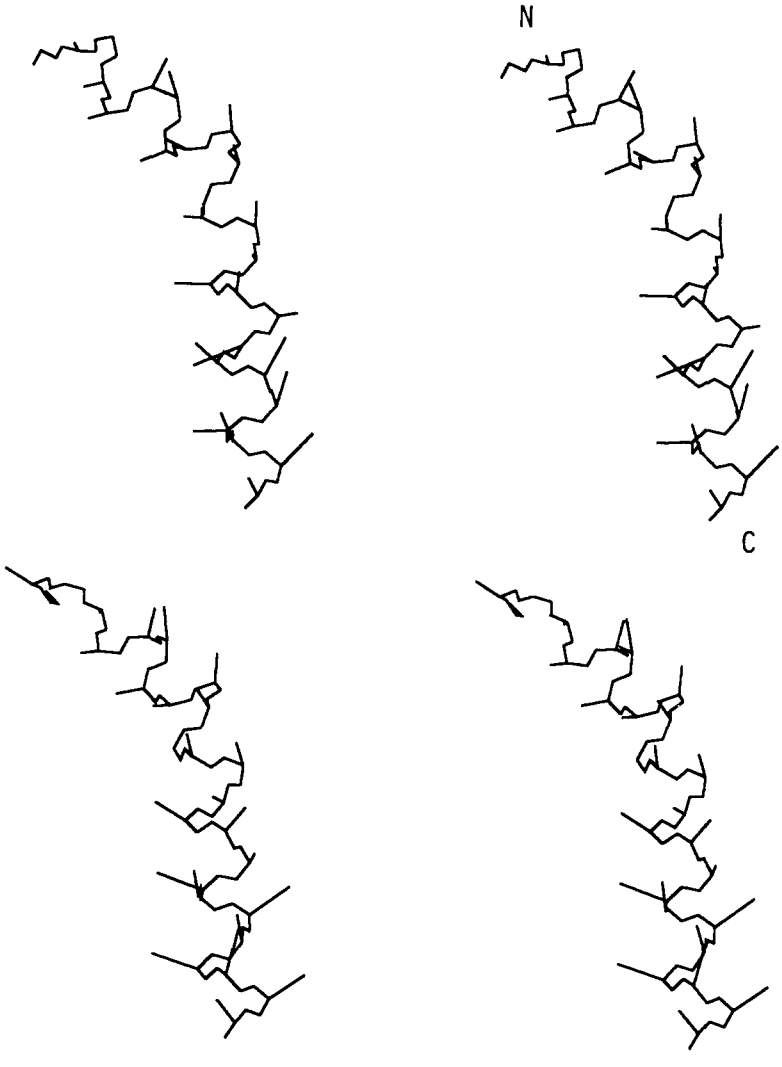

Fig. 3. Backbone and side-chain centroid stereo plot for melittin. **Top:** The crystal structure. **Bottom:** One of the 90 computed structures (RMS = 1.64 Å, DME = 0.98 Å to the crystal structure). All the other computed structures are similar to this one with RMS less than 0.15 Å (compared to this computed structure).

which the heterogeneity among the amino acids due to the details of the side-chain interactions has been partially eliminated.

Pincus et al. (1982) computed folded conformations of the 20 N-terminal residues of melittin by a buildup procedure. Their method uses limited numbers of locally optimized di- and tripeptide conformations to build up progressively global conformations from N-terminal to C-terminal. The genetic algorithm RRM described here is very different from that of Pincus et al.

APPI

Similar results have been found for APPI. Table 4 lists the simulation results for APPI. The primary sequence and the radius of gyration from the APPI crystal structure were the only input information for the simulation. The penalty coefficient λ was set to be 280 units/Å. The random perturbation parameter d was the same as that in the simulation for melittin. Convergence in the energy profile and in the three-dimensional conformation has been found to be similar to that in the melittin simulation. Figure 4 shows the energy profile (blank bar) for all 90 initial conformations and the energy profile (filled bar) for the conformation after the minimization. The RMS between any two computed structures is less than 0.30 Å. While the average radius of gyration of 90 computed structures, 10.6 Å, is fairly close to the value for crystal structure of 10.7 Å, the average total contacts of the computed structures is 466.5, which is significantly different from the crystal value of 530. This deviation primarily arises from the different folding of five C-terminal residues (see below) compared to their crystallographic positions. The computed structures folded less tightly than the crystal structure does. Figure 5 shows the stereo plot of backbone and side-chain centroids for APPI crystal and one of the computed structures. The DME and the RMS to crystal structure for the computed structures are not as good as the corresponding values of melittin. I have noticed, however, that this large RMS value is primarily due to the fact that the five C-terminal end residues of APPI in the computed structures fold significantly differ-

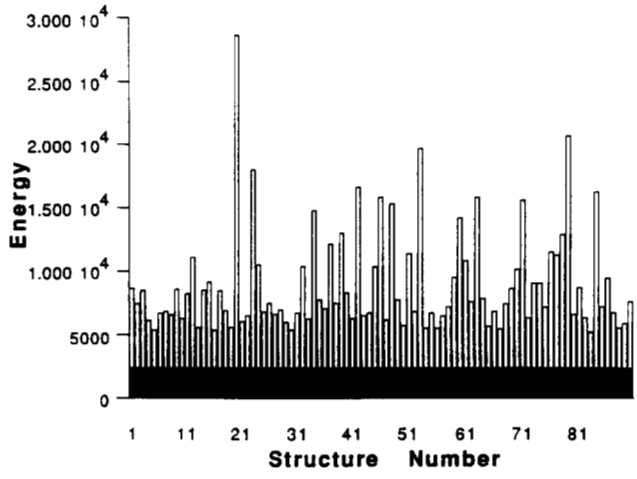

Fig. 4. Energy profile in APPI simulation. Energy profile of the randomly created initial conformations (open bars) and the energy profile for the conformation after the RRM genetic algorithm minimization (filled bars). The average energy of the initial conformations is 6,649.35 units with a standard deviation of 4,089.66 units, and the average energy for the conformation after the minimization is 2,379.15 units with a standard deviation of 1.36 units.

ently from the last five residues in the crystal structure. RMS† in Table 4 lists the RMS of the computed conformations to the crystal structure when the five C-terminal end residues are not included in the structural alignment. The RMS† values are substantially smaller than the corresponding values when the whole computed structure is aligned with the crystal structure. In other words, 31 residues out of 36 residues in APPI fold correctly in the current RRM model with an accuracy of average RMS of 1.30 Å to the crystal structure. It is very interesting to note that the C-terminal end of APPI is relatively flexible even in the crystal structure (Glover et al., 1983).

Apamin

Apamin is an 18-residue polypeptide component of bee venom. Native apamin contains two disulfide bonds linked between residues 1 and 11 and between 3 and 15. In the simulations for apamin, the segmentation probabilities have been set to $P_2 = 0.6$, $P_3 = 0.4$, and $P_4 = P_5 = 0.0$. Random perturbation of Equation 16 was used with

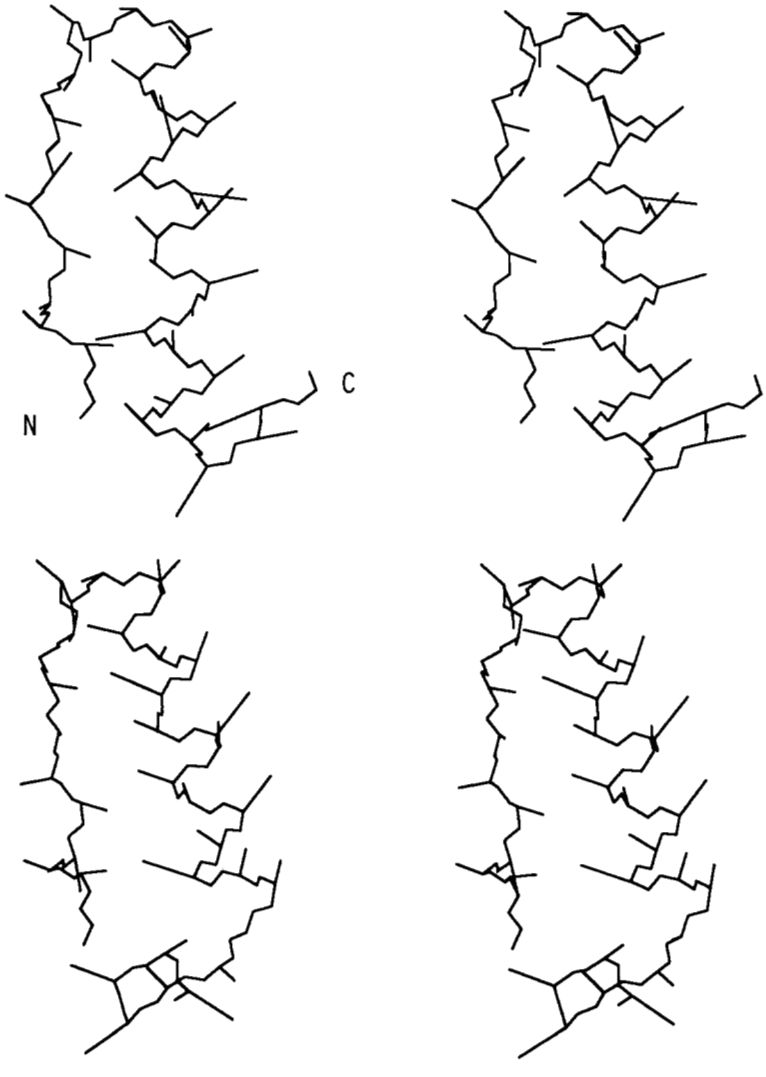

Fig. 5. Backbone and side-chain centroid stereo plots for avian pancreatic polypeptide inhibitor. **Top:** The crystal structure. **Bottom:** One of the 90 computed structures (RMS = 3.93 Å, DME = 1.75 Å to the crystal structure). If only 31 out of 36 residues are structurally aligned with the crystal structure, the RMS = 1.29 Å, and DME = 1.05 Å. All the other computed structures are similar to this one with RMS less than 0.30 Å.

Table 4. *Simulations for avian pancreatic polypeptide inhibitor, a protein of 36 residues*[a]

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	RMS†	Total contacts	Radius of gyration
1	6,300.85	2,377.02	673.02	729.16	430.24	541.88	1.76	3.94	1.32	466	10.6
2	5,105.48	2,377.02	673.02	729.16	430.24	541.88	1.76	3.94	1.32	466	10.6
3	6,116.62	2,377.02	673.02	729.16	430.24	541.88	1.76	3.94	1.32	466	10.6
4	3,769.54	2,377.10	673.21	729.63	429.85	541.88	1.75	3.93	1.29	468	10.6
5	3,037.40	2,377.10	673.21	729.63	429.85	541.88	1.75	3.93	1.29	468	10.6
6	4,369.94	2,377.10	673.21	729.63	429.85	541.88	1.75	3.93	1.29	468	10.6
7	4,522.96	2,377.33	673.19	729.16	430.24	541.88	1.76	3.94	1.33	466	10.6
8	4,220.98	2,377.33	673.19	729.16	430.24	541.88	1.76	3.94	1.33	466	10.6
9	6,244.32	2,377.33	673.19	729.16	430.24	541.88	1.76	3.94	1.33	466	10.6
10	3,945.14	2,377.46	673.02	729.16	430.68	541.88	1.76	3.94	1.32	466	10.6
11	5,890.29	2,377.46	673.02	729.16	430.68	541.88	1.76	3.94	1.32	466	10.6
12	8,773.48	2,377.46	673.02	729.16	430.68	541.88	1.76	3.94	1.32	466	10.6
13	3,228.33	2,377.54	673.21	729.63	430.29	541.88	1.75	3.93	1.29	468	10.6
14	6,161.75	2,377.54	673.21	729.63	430.29	541.88	1.75	3.93	1.29	468	10.6
15	6,831.68	2,377.54	673.21	729.63	430.29	541.88	1.75	3.93	1.29	468	10.6
16	3,027.06	2,377.77	673.19	729.16	430.68	541.88	1.76	3.94	1.33	466	10.6
17	6,129.39	2,377.77	673.19	729.16	430.68	541.88	1.76	3.94	1.33	466	10.6
18	4,596.85	2,377.77	673.19	729.16	430.68	541.88	1.76	3.94	1.33	466	10.6
19	3,245.19	2,377.92	673.04	730.08	430.24	541.88	1.76	3.94	1.33	464	10.6
20	26,335.22	2,377.92	673.04	730.08	430.24	541.88	1.76	3.94	1.33	464	10.6
21	3,714.76	2,377.92	673.04	730.08	430.24	541.88	1.76	3.94	1.33	464	10.6
22	4,183.21	2,377.95	673.90	730.06	429.02	541.88	1.76	3.93	1.32	466	10.6
23	15,643.95	2,377.95	673.90	730.06	429.02	541.88	1.76	3.93	1.32	466	10.6
24	8,180.30	2,377.95	673.90	730.06	429.02	541.88	1.76	3.93	1.32	466	10.6
25	4,477.98	2,378.13	673.72	729.85	429.42	541.88	1.77	3.94	1.36	460	10.7
26	5,143.52	2,378.13	673.72	729.85	429.42	541.88	1.77	3.94	1.36	460	10.7
27	4,289.66	2,378.13	673.72	729.85	429.42	541.88	1.77	3.94	1.36	460	10.7
28	4,661.05	2,378.36	673.04	730.08	430.68	541.88	1.76	3.94	1.33	464	10.6
29	3,648.62	2,378.36	673.04	730.08	430.68	541.88	1.76	3.94	1.33	464	10.6
30	3,050.06	2,378.36	673.04	730.08	430.68	541.88	1.76	3.94	1.33	464	10.6
31	4,373.98	2,378.39	673.90	730.06	429.46	541.88	1.76	3.93	1.32	466	10.6
32	8,084.94	2,378.39	673.90	730.06	429.46	541.88	1.76	3.93	1.32	466	10.6
33	3,920.38	2,378.39	673.90	730.06	429.46	541.88	1.76	3.93	1.32	466	10.6
34	12,452.85	2,378.47	674.05	729.29	430.56	541.88	1.77	3.95	1.38	464	10.6
35	5,401.57	2,378.47	674.05	729.29	430.56	541.88	1.77	3.95	1.38	464	10.6
36	4,744.02	2,378.47	674.05	729.29	430.56	541.88	1.77	3.95	1.38	464	10.6
37	9,813.67	2,378.49	673.72	730.17	429.42	541.88	1.77	3.94	1.35	460	10.7
38	5,159.90	2,378.49	673.72	730.17	429.42	541.88	1.77	3.94	1.35	460	10.7
39	10,683.21	2,378.49	673.72	730.17	429.42	541.88	1.77	3.94	1.35	460	10.7
40	5,958.69	2,378.57	673.72	729.85	429.85	541.88	1.77	3.94	1.36	460	10.7
41	3,935.60	2,378.57	673.72	729.85	429.85	541.88	1.77	3.94	1.36	460	10.7
42	14,258.66	2,378.57	673.72	729.85	429.85	541.88	1.77	3.94	1.36	460	10.7
43	4,185.75	2,378.91	674.05	729.29	431.00	541.88	1.77	3.95	1.38	464	10.6
44	4,394.35	2,378.91	674.05	729.29	431.00	541.88	1.77	3.95	1.38	464	10.6
45	8,043.25	2,378.91	674.05	729.29	431.00	541.88	1.77	3.95	1.38	464	10.6
46	13,461.39	2,378.93	673.72	730.17	429.85	541.88	1.77	3.94	1.35	460	10.7
47	3,819.72	2,378.93	673.72	730.17	429.85	541.88	1.77	3.94	1.35	460	10.7
48	12,971.94	2,378.93	673.72	730.17	429.85	541.88	1.77	3.94	1.35	460	10.7
49	5,401.65	2,379.05	674.07	730.23	429.42	541.88	1.77	3.94	1.36	460	10.7
50	3,411.03	2,379.05	674.07	730.23	429.42	541.88	1.77	3.94	1.36	460	10.7
51	9,080.26	2,379.05	674.07	730.23	429.42	541.88	1.77	3.94	1.36	460	10.7
52	4,507.33	2,379.49	674.07	730.23	429.85	541.88	1.77	3.94	1.36	460	10.7
53	17,334.40	2,379.49	674.07	730.23	429.85	541.88	1.77	3.94	1.36	460	10.7
54	3,186.79	2,379.49	674.07	730.23	429.85	541.88	1.77	3.94	1.36	460	10.7
55	4,385.97	2,379.67	676.71	732.16	428.74	541.88	1.77	3.99	1.18	478	10.6
56	3,191.72	2,379.67	676.71	732.16	428.74	541.88	1.77	3.99	1.18	478	10.6
57	4,161.78	2,379.67	676.71	732.16	428.74	541.88	1.77	3.99	1.18	478	10.6
58	4,896.73	2,380.09	675.76	729.42	429.74	541.88	1.78	3.95	1.41	460	10.7
59	7,200.89	2,380.09	675.76	729.42	429.74	541.88	1.78	3.95	1.41	460	10.7
60	11,864.37	2,380.09	675.76	729.42	429.74	541.88	1.78	3.95	1.41	460	10.7
61	8,503.88	2,380.11	676.71	732.16	429.18	541.88	1.77	3.99	1.18	478	10.6

(*continued*)

Table 4. *Continued*

No.	E_{start}	E_{end}	E_{C^α}	E_{SC}	E_S	E_D	DME	RMS	RMS†	Total contacts	Radius of gyration
62	5,261.20	2,380.11	676.71	732.16	429.18	541.88	1.77	3.99	1.18	478	10.6
63	13,482.49	2,380.11	676.71	732.16	429.18	541.88	1.77	3.99	1.18	478	10.6
64	5,500.17	2,380.53	675.76	729.42	430.18	541.88	1.78	3.95	1.41	460	10.7
65	3,305.59	2,380.53	675.76	729.42	430.18	541.88	1.78	3.95	1.41	460	10.7
66	4,490.90	2,380.53	675.76	729.42	430.18	541.88	1.78	3.95	1.41	460	10.7
67	3,106.92	2,380.58	677.12	731.93	429.46	541.88	1.78	4.00	1.27	470	10.6
68	5,107.46	2,380.58	677.12	731.93	429.46	541.88	1.78	4.00	1.27	470	10.6
69	6,298.07	2,380.58	677.12	731.93	429.46	541.88	1.78	4.00	1.27	470	10.6
70	7,815.83	2,380.63	676.76	733.74	427.92	541.88	1.77	3.98	1.21	472	10.6
71	13,242.62	2,380.63	676.76	733.74	427.92	541.88	1.77	3.98	1.21	472	10.6
72	3,991.96	2,380.63	676.76	733.74	427.92	541.88	1.77	3.98	1.21	472	10.6
73	6,741.78	2,380.82	677.06	732.54	429.13	541.88	1.77	3.99	1.22	474	10.6
74	6,736.16	2,380.82	677.06	732.54	429.13	541.88	1.77	3.99	1.22	474	10.6
75	4,821.90	2,380.82	677.06	732.54	429.13	541.88	1.77	3.99	1.22	474	10.6
76	9,172.04	2,380.98	676.78	732.97	429.13	541.88	1.77	3.99	1.21	474	10.6
77	8,960.64	2,380.98	676.78	732.97	429.13	541.88	1.77	3.99	1.21	474	10.6
78	10,543.74	2,380.98	676.78	732.97	429.13	541.88	1.77	3.99	1.21	474	10.6
79	18,306.99	2,381.02	676.35	734.10	428.31	541.88	1.78	3.99	1.25	470	10.6
80	4,277.75	2,381.02	676.35	734.10	428.31	541.88	1.78	3.99	1.25	470	10.6
81	6,389.26	2,381.02	676.35	734.10	428.31	541.88	1.78	3.99	1.25	470	10.6
82	4,004.19	2,381.02	677.12	731.93	429.90	541.88	1.78	4.00	1.27	470	10.6
83	2,871.08	2,381.02	677.12	731.93	429.90	541.88	1.78	4.00	1.27	470	10.6
84	13,896.76	2,381.02	677.12	731.93	429.90	541.88	1.78	4.00	1.27	470	10.6
85	4,873.84	2,381.07	676.76	733.74	428.35	541.88	1.77	3.98	1.21	472	10.6
86	7,121.87	2,381.07	676.76	733.74	428.35	541.88	1.77	3.98	1.21	472	10.6
87	4,409.48	2,381.07	676.76	733.74	428.35	541.88	1.77	3.98	1.21	472	10.6
88	3,217.95	2,381.11	676.61	733.91	428.31	541.88	1.78	3.99	1.24	470	10.6
89	3,560.30	2,381.11	676.61	733.91	428.31	541.88	1.78	3.99	1.24	470	10.6
90	5,296.05	2,381.11	676.61	733.91	428.31	541.88	1.78	3.99	1.24	470	10.6
Average	6,649.35	2,379.15	674.78	730.78	429.64	541.88	1.77	3.96	1.30	466.5	10.6
σ	4,089.66	1.36	1.57	1.63	0.79	0.00	0.01	0.02	0.06	5.4	0.0

[a] The notations used in the table are the same as in Table 2. RMS† denotes the RMS to the crystal structure when five residues in the C-terminal end are not included in the structure alignment. Starting conformations were randomly created. The crystal structure has 530 total contacts and a radius of gyration of 10.68 Å.

$d = 10°$. The simulation input is the primary sequence and the radius of gyration of the DISCOVER-minimized NMR structure of apamin. Two sets of simulations have been carried out: (case 1) no further information other than the above; (case 2) explicit disulfide bond pairings also were given, and this information was implemented by an extra energy term. The nonlocal interaction potential between the cysteines that form disulfide bonds was set to be twice as large as that for those cysteines that are not explicitly noted to form disulfide bonds. In other words, the nonlocal interaction potential was doubled for Cys-1–Cys-11 and Cys-3–Cys-15.

Results of the simulations are listed in Table 5. The total energy profile of the starting conformations and the final conformations has the same characteristics as that found in the simulation of melittin and APPI for both cases 1 and 2 above. The final structures, in case 1, have an average DME and an average RMS of 2.53 Å and 3.38 Å, respectively. In these final structures 15 out of 90 have an RMS less than 3.0 Å. The average total contacts and the average radius of gyration of these 90 optimized structures are 239 and 6.2 Å, respectively, figures very close to the corresponding values of 238 and 6.4 Å for the DISCOVER-minimized NMR structure. The final structures, in case 2, have an average DME and an average RMS of 2.24 Å and 2.79 Å. The average total contacts and the average gyration radius are 243 and 6.0 Å, which indicates that the computed structures are slightly more compact than the NMR structure. With the explicit information of disulfide bond pairings the computed structures have improved the average RMS by about 0.6 Å. The computed structures in both cases have a very high similarity among themselves. The RMS between any two computed structures is less than 1.5 Å (most of them are less than 0.5 Å) among the computed structures in case 1 and less than 0.3 Å among the computed structures in case 2. Figure 6 gives the stereo plots of the apamin DISCOVER-minimized NMR structure, one of the com-

Table 5. *Simulations for apamin, a protein of 18 residues with two disulfide bonds*[a]

No.	E_{start} (a)	E_{end} (a)	E_{start} (b)	E_{end} (b)	DME (a,b)		RMS (a,b)		Total contacts (a,b)		Radius of gyration (a,b)	
1	4,608.25	676.64	4,940.03	647.29	2.54	2.23	3.46	2.80	236	242	6.3	6.0
2	1,697.81	676.64	1,752.76	647.29	2.54	2.23	3.46	2.80	236	242	6.3	6.0
3	4,727.07	676.64	5,066.58	647.29	2.54	2.23	3.46	2.80	236	242	6.3	6.0
4	1,132.31	677.86	1,146.40	648.13	2.53	2.24	3.45	2.81	236	242	6.3	6.0
5	2,934.85	677.86	2,975.88	648.13	2.53	2.24	3.45	2.81	236	242	6.3	6.0
6	1,287.35	677.86	1,283.65	648.13	2.53	2.24	3.45	2.81	236	242	6.3	6.0
7	1,556.87	679.63	1,558.01	648.45	2.54	2.23	3.46	2.80	236	242	6.3	6.1
8	1,031.37	679.63	1,040.33	648.45	2.54	2.23	3.46	2.80	236	242	6.3	6.1
9	1,795.32	679.63	1,876.56	648.45	2.54	2.23	3.46	2.80	236	242	6.3	6.1
10	2,074.14	680.30	2,142.67	652.39	2.55	2.23	3.48	2.78	238	244	6.3	6.0
11	3,238.85	680.30	3,451.78	652.39	2.55	2.23	3.48	2.78	238	244	6.3	6.0
12	2,988.66	680.30	3,109.11	652.39	2.55	2.23	3.48	2.78	238	244	6.3	6.0
13	2,204.34	680.85	2,246.52	653.44	2.53	2.23	3.46	2.79	236	244	6.3	6.0
14	918.79	680.85	920.34	653.44	2.53	2.23	3.46	2.79	236	244	6.3	6.0
15	3,984.66	680.85	4,261.62	653.44	2.53	2.23	3.46	2.79	236	244	6.3	6.0
16	876.89	680.96	879.44	653.46	2.55	2.24	3.48	2.79	238	242	6.3	6.0
17	3,377.51	680.96	3,588.34	653.46	2.55	2.24	3.48	2.79	238	242	6.3	6.0
18	1,726.38	680.96	1,725.20	653.46	2.55	2.24	3.48	2.79	238	242	6.3	6.0
19	939.84	681.30	949.68	654.41	2.30	2.22	2.90	2.79	248	242	6.1	6.1
20	3,184.39	681.30	3,383.50	654.41	2.30	2.22	2.90	2.79	248	242	6.1	6.1
21	977.02	681.30	978.72	654.41	2.30	2.22	2.90	2.79	248	242	6.1	6.1
22	1,058.65	681.97	1,068.04	655.07	2.54	2.23	3.46	2.80	236	242	6.3	6.0
23	2,060.97	681.97	2,152.48	655.07	2.54	2.23	3.46	2.80	236	242	6.3	6.0
24	3,797.78	681.97	4,051.83	655.07	2.54	2.23	3.46	2.80	236	242	6.3	6.0
25	2,670.74	682.14	2,818.47	655.11	2.64	2.23	3.56	2.78	234	244	6.3	6.0
26	1,545.83	682.14	1,566.69	655.11	2.64	2.23	3.56	2.78	234	244	6.3	6.0
27	3,477.23	682.14	3,702.55	655.11	2.64	2.23	3.56	2.78	234	244	6.3	6.0
28	3,841.50	682.64	4,099.76	655.20	2.54	2.24	3.46	2.78	236	242	6.3	6.0
29	2,223.20	682.64	2,245.55	655.20	2.54	2.24	3.46	2.78	236	242	6.3	6.0
30	1,603.14	682.64	1,661.60	655.20	2.54	2.24	3.46	2.78	236	242	6.3	6.0
31	1,834.57	682.65	1,854.16	655.88	2.57	2.23	3.40	2.77	242	244	6.1	6.0
32	4,165.77	682.65	4,452.07	655.88	2.57	2.23	3.40	2.77	242	244	6.1	6.0
33	6,080.29	682.65	6,540.58	655.88	2.57	2.23	3.40	2.77	242	244	6.1	6.0
34	2,171.25	682.76	2,289.44	655.89	2.33	2.24	2.93	2.80	250	242	6.0	6.0
35	1,531.48	682.76	1,589.83	655.89	2.33	2.24	2.93	2.80	250	242	6.0	6.0
36	1,398.34	682.76	1,420.17	655.89	2.33	2.24	2.93	2.80	250	242	6.0	6.0
37	2,581.97	682.80	2,719.58	656.17	2.65	2.24	3.56	2.79	234	242	6.3	6.0
38	2,823.36	682.80	2,990.07	656.17	2.65	2.24	3.56	2.79	234	242	6.3	6.0
39	890.48	682.80	894.15	656.17	2.65	2.24	3.56	2.79	234	242	6.3	6.0
40	1,026.29	682.83	1,032.90	657.35	2.54	2.23	3.46	2.77	236	244	6.3	6.0
41	2,142.13	682.83	2,246.77	657.35	2.54	2.23	3.46	2.77	236	244	6.3	6.0
42	3,223.83	682.83	3,423.03	657.35	2.54	2.23	3.46	2.77	236	244	6.3	6.0
43	3,451.35	683.09	3,453.77	657.50	2.29	2.25	2.90	2.80	248	244	6.1	6.0
44	1,484.12	683.09	1,534.62	657.50	2.29	2.25	2.90	2.80	248	244	6.1	6.0
45	3,414.89	683.09	3,628.69	657.50	2.29	2.25	2.90	2.80	248	244	6.1	6.0
46	1,378.15	683.26	1,418.96	658.05	2.57	2.24	3.40	2.79	242	244	6.1	6.0
47	2,410.50	683.26	2,534.86	658.05	2.57	2.24	3.40	2.79	242	244	6.1	6.0
48	1,397.08	683.26	1,438.01	658.05	2.57	2.24	3.40	2.79	242	244	6.1	6.0
49	1,913.48	683.39	2,001.18	658.35	2.30	2.25	2.91	2.79	248	244	6.1	6.0
50	2,413.79	683.39	2,545.71	658.35	2.30	2.25	2.91	2.79	248	244	6.1	6.0
51	2,555.90	683.39	2,703.10	658.35	2.30	2.25	2.91	2.79	248	244	6.1	6.0
52	1,981.59	683.67	1,981.15	658.76	2.57	2.26	3.40	2.81	242	244	6.1	6.0
53	1,695.46	683.67	1,772.84	658.76	2.57	2.26	3.40	2.81	242	244	6.1	6.0
54	3,929.91	683.67	4,189.59	658.76	2.57	2.26	3.40	2.81	242	244	6.1	6.0
55	1,584.87	683.83	1,642.70	659.02	2.65	2.26	3.56	2.81	234	242	6.2	6.0
56	2,627.37	683.83	2,778.94	659.02	2.65	2.26	3.56	2.81	234	242	6.2	6.0
57	1,466.07	683.83	1,506.71	659.02	2.65	2.26	3.56	2.81	234	242	6.2	6.0
58	1,353.28	683.96	1,396.41	659.15	2.64	2.23	3.55	2.77	234	244	6.3	6.0
59	1,935.34	683.96	2,028.00	659.15	2.64	2.23	3.55	2.77	234	244	6.3	6.0
60	1,000.23	683.96	1,014.53	659.15	2.64	2.23	3.55	2.77	234	244	6.3	6.0

(*continued*)

Table 5. *Continued*

No.	E_{start} (a)	E_{end} (a)	E_{start} (b)	E_{end} (b)	DME (a,b)		RMS (a,b)		Total contacts (a,b)		Radius of gyration (a,b)	
61	2,706.64	684.49	2,867.47	659.36	2.65	2.25	3.56	2.80	234	242	6.2	6.0
62	1,092.95	684.49	1,106.74	659.36	2.65	2.25	3.56	2.80	234	242	6.2	6.0
63	1,307.17	684.49	1,311.26	659.36	2.65	2.25	3.56	2.80	234	242	6.2	6.0
64	2,079.12	684.61	2,179.85	660.21	2.53	2.24	3.45	2.81	236	242	6.3	6.0
65	907.68	684.61	907.40	660.21	2.53	2.24	3.45	2.81	236	242	6.3	6.0
66	917.00	684.61	919.85	660.21	2.53	2.24	3.45	2.81	236	242	6.3	6.0
67	1,011.21	684.62	1,026.52	660.23	2.64	2.23	3.55	2.80	234	242	6.3	6.0
68	985.76	684.62	990.47	660.23	2.64	2.23	3.55	2.80	234	242	6.3	6.0
69	1,317.61	684.62	1,357.24	660.23	2.64	2.23	3.55	2.80	234	242	6.3	6.0
70	2,735.13	685.38	2,901.97	661.54	2.57	2.24	3.40	2.79	242	242	6.1	6.0
71	804.42	685.38	804.05	661.54	2.57	2.24	3.40	2.79	242	242	6.1	6.0
72	2,280.51	685.38	2,412.07	661.54	2.57	2.24	3.40	2.79	242	242	6.1	6.0
73	2,931.12	685.70	3,113.70	661.61	2.35	2.24	2.94	2.80	234	242	6.2	6.0
74	2,028.29	685.70	2,090.90	661.61	2.35	2.24	2.94	2.80	234	242	6.2	6.0
75	5,820.01	685.70	6,254.17	661.61	2.35	2.24	2.94	2.80	234	242	6.2	6.0
76	4,928.56	685.81	5,291.35	661.92	2.54	2.26	3.46	2.81	236	242	6.3	6.0
77	1,938.13	685.81	2,016.91	661.92	2.54	2.26	3.46	2.81	236	242	6.3	6.0
78	2,044.68	685.81	2,049.63	661.92	2.54	2.26	3.46	2.81	236	242	6.3	6.0
79	1,282.42	685.97	1,307.65	662.70	2.56	2.27	3.38	2.82	242	242	6.2	6.0
80	3,323.07	685.97	3,535.32	662.70	2.56	2.27	3.38	2.82	242	242	6.2	6.0
81	1,439.44	685.97	1,494.59	662.70	2.56	2.27	3.38	2.82	242	242	6.2	6.0
82	1,898.57	686.17	1,933.88	663.23	2.55	2.25	3.48	2.79	238	244	6.3	6.0
83	977.66	686.17	990.61	663.23	2.55	2.25	3.48	2.79	238	244	6.3	6.0
84	1,283.22	686.17	1,166.72	663.23	2.55	2.25	3.48	2.79	238	244	6.3	6.0
85	3,770.57	686.39	3,732.30	663.29	2.56	2.26	3.38	2.81	242	244	6.2	6.0
86	1,636.43	686.39	1,697.23	663.29	2.56	2.26	3.38	2.81	242	244	6.2	6.0
87	2,302.26	686.39	2,430.61	663.29	2.56	2.26	3.38	2.81	242	244	6.2	6.0
88	2,047.03	686.84	2,066.36	663.47	2.55	2.23	3.48	2.78	238	244	6.3	6.0
89	2,332.90	686.84	2,462.03	663.47	2.55	2.23	3.48	2.78	238	244	6.3	6.0
90	3,792.49	686.84	4,047.82	663.47	2.55	2.23	3.48	2.78	238	244	6.3	6.0
Average	2,236.97	683.08	2,334.81	657.22	2.53	2.24	3.38	2.79	239	243	6.2	6.0
σ	1,159.99	2.43	1,259.34	4.39	0.10	0.01	0.21	0.01	4.8	1.0	0.1	0.0

[a] Two sets of data are presented in the table: a: the simulation without constraints on disulfide bonds, and b: the simulation with the explicit native disulfide bond constraints, which was implemented by an additional nonlocal interaction potential for cysteines that form disulfide bonds. The segmentation probabilities used in the simulations are $P_2 = 0.6$, $P_3 = 0.4$, and $P_4 = P_5 = 0.0$. The reference structure used in the calculation of DME and RMS of the optimized structures was a DISCOVER-minimized NMR structure of apamin. This NMR-measured structure has a radius of gyration of 6.4 Å, and 238 total contacts. All the initial conformations were created randomly. The penalty coefficient λ was set to 120 units/Å.

puted structures in case 1, and one of the computed structures in case 2.

One of the very interesting features in the computed structures in case 1 is that distances between C^α atom of Cys-3 and Cys-15 in different optimized structures are very close, ranging from 4.95 Å to 5.79 Å. The average distance between C^α atom of Cys-3 and Cys-15 for 90 computed structures is 5.55 Å, which is an optimal C^α-C^α distance to form a disulfide bond (Srinivasan et al., 1990). The average distance between cysteines 1 and 3 is 5.70 Å (ranging from 5.32 Å to 6.90 Å), 7.58 Å between cysteines 11 and 15, 10.25 Å between cysteines 3 and 11, and 9.79 Å between cysteines 1 and 15. On the other hand, for Cys-1 and Cys-11, only 15 out of 90 computed structures have a C^α-C^α distance that may be favorable for disulfide bond formation (the average is 13.22 Å). These 15 computed structures have an RMS less than 3.0 Å. These results were obtained without the information of disulfide bond pairing in the simulation, and they suggest that the native disulfide bond pairings are between cysteines 3 and 15 and 1 and 11. They also suggest that the disulfide bond between cysteines 3 and 15 is formed with higher probability than that between cysteines 1 and 11 as an experiment has indicated (Huyghues-Despointes & Nelson, 1992).

When the disulfide bond information is included as an additional nonlocal interaction potential between the cysteines that form disulfide bond, the computed structures are more native-like (case 2). The C^α-C^α distance between Cys-1 and Cys-11 is 6.85 Å and 5.67 Å between Cys-3 and Cys-15.

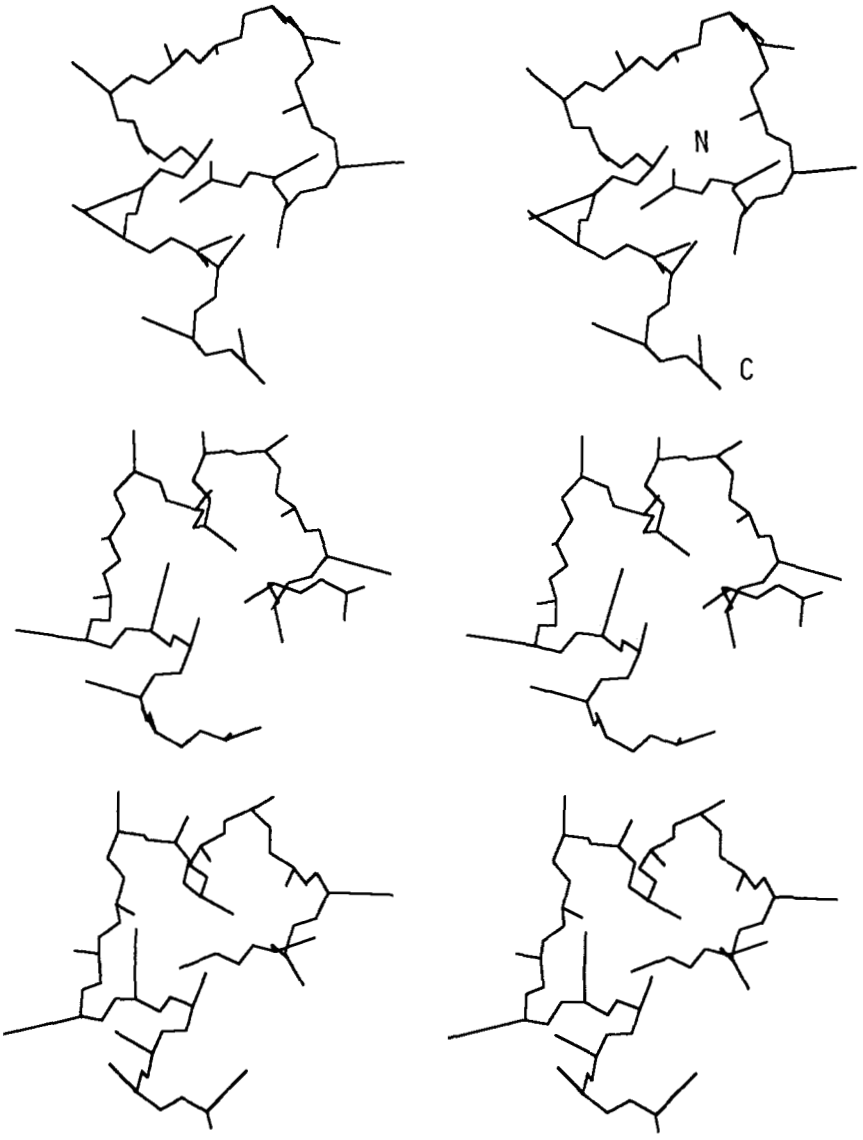

Fig. 6. Backbone and side-chain centroid stereo plots for apamin. **Top:** The DISCOVER-minimized NMR structure. **Middle:** One of the 90 computed structures (DME = 2.54 Å, RMS = 3.46 Å) without explicit disulfide bond information in the simulation. **Bottom:** One of the 90 computed structures (DME = 2.23 Å, RMS = 2.77 Å) with the explicit disulfide bond information in the simulation.

Effect of random perturbation and E_{rg}

A random perturbation term has been added in the change of each geometric variable (ϕ, ψ) in the dictionary-assisted conformational search process (see Equation 16). In doing so, it is hoped that this mechanism will be able to explore a much larger region in the conformational space than the possible combinations of the segmental conformations from the primary conformation dictionaries for a given primary sequence, so that the search may be more complete.

Previous studies (Sun & Luo, in prep.) indicated that the statistical potential function used in the current reduced representation model may not only deform the mean energy surface of the system but may also introduce further degeneracy, which corresponds to several different structures of comparable energy. There are several possible ways to improve the simple statistical potential function on the empirical level (or adding phenomenological interaction terms), as mentioned in section II. In this study, a penalty energy term for radius of gyration, Equation 10 has been added to the statistical potential function.

Table 6 listed the average results of 90 structures computed under different conditions as indicated in the first column of the table. The results show that with the combined application of the random perturbation and the target function for radius of gyration, the computed structures are much closer to the corresponding crystal structures.

Efficiency of algorithm

The RRM genetic algorithm minimization method presented here is very robust and has a very high degree of convergence. Three important factors greatly improve the

Table 6. *Effect of the random pertubation and* E_{rg}[a]

Structure	Average total contacts	Average radius of gyration (Å)	Average DME (Å)	Average RMS (Å)
Melittin crystal	300.0	11.1		
R-T	304.9	10.8	0.99	1.66
NR-T	284.9	10.8	1.82	2.67
R-NT	319.9	8.7	4.81	5.78
NR-NT	306.5	9.0	4.95	5.59
APPI crystal	530.0	10.7		
R-T	466.5	10.6	1.77	3.96
NR-T	450.1	10.4	3.29	5.15
NR-NT	430.9	10.7	2.79	4.64

[a] All the listed average data are for 90 structures. Melittin crystal and APPI crystal are the crystal structures, all the rest are average values of the computed structures. R and NR indicate including and not including the random perturbation in Equation 16, respectively. T and NT indicates including and not including the energy term of Equation 10.

efficiency of the computation: (1) the intrinsic parallel search of the genetic algorithm, which provides a simultaneous multipath search algorithm in the conformational space; (2) the use of the dictionary-assisted variation of the conformations, which excludes all possible conformations that have local steric conflicts; and (3) the use of a penalty energy function (or target function) Equation 10, which selects conformations with a preferred value of radius of gyration. In general, the number of structures that have been simultaneously computed are $5p$, where p is the size of the conformational population. A 90-structure population will have had 450 structures processed simultaneously. In the case of the melittin simulation (population size 90), the minimization converged after about 32–45 generations of the genetic algorithm search operations in a total central processing unit (CPU) time of 5 min on a VAX-6400. For APPI simulation, the minimization converged after about 37–55 generations of genetic algorithm search operations in a total CPU time of about 10 min on a VAX-6400. In another report (Sun & Luo, in prep.), we used a simulated annealing minimization algorithm and computed the optimized structures for melittin and APPI in a similar reduced representation model. It took about 10 min to optimize one melittin structure and about 15 min for one APPI structure. The current RRM genetic algorithms minimization algorithm is roughly 100–200 times faster than the RRM simulated annealing algorithm. Moreover, the current RRM genetic algorithm is able to generate much more convergent and accurate conformations than that of the RRM simulated annealing algorithm, at least in the cases of melittin and APPI. Extensive simulated annealing calculation for apamin has been carried out (Sun & Snyder, in prep.), the results show that both the secondary structure element and the disulfide bridges are formed naturally. However, the simulated annealing computation is slower by a factor of 200 in comparison with the computation by genetic algorithms.

Discussion

The reduced representation model and dictionary-assisted genetic algorithms conformational search method described in this paper have been tested on three small proteins, melittin, APPI, and apamin. The results have shown that the computed conformations are convergent in both their energy profiles and three-dimensional structures. With the primary sequence and the radius of gyration of the crystal structure as the only input information for the computation, the computed structures are close to their corresponding crystal structures (for melittin and APPI) or NMR structure (apamin). This indicates that the current model is consistent with the basic characteristics of the folded protein structures.

The performance of the statistical potential function computed from the known structures of proteins and used in the current model has been further improved by the additional phenomenological energy term, i.e., E_{rg} the target energy function for radius of gyration. This energy term uses the specific crystal structure information, radius of gyration R_g, of the protein to be computed. I consider that this constraint of the radius of gyration is a low-order one, because the number of possible folded conformations is still enormously large if the radius of gyration is the only folding constraint. Without this term, the computed structures still fold into ones that have overall similarity to the crystal structures (Sun & Luo, in prep.). Instead of the constraint for the radius of gyration, there are a number of other possible ways to improve the performance of the statistical potential function further. For instance, the statistical radial distribution functions of amino acids in the known proteins (Sun, in prep.) could be used for this purpose.

The use of short peptide conformations as found in known protein structures as the primary conformation pool to search protein conformations has been discussed by several authors, these studies, however, have focused on the use of hexapeptide conformations (e.g., Unger et al., 1989). For a given protein sequence, there are not many samples for hexapeptide conformations in the known protein structures, and if similar sequences are eliminated, the sample size is even smaller. This may result in a serious limitation in sampling conformational space. The dictionary-assisted conformational search method presented here attempts to overcome come this problem by using a set of shorter peptide conformational pools. While the dictionary-assisted conformational search reduces the number of conformations to be searched by a factor enormously large in comparison to the random search, the variability of the accessible conformations is retained by the use of the shorter segmental conforma-

tions and their random perturbations. The (ϕ, ψ) sampling probability is automatically in accordance with the Ramachandran distribution of (ϕ, ψ) for any amino acid residue in the segmental method; furthermore, higher-order correlation in short sequence segments is utilized.

Careful readers would have noticed that the 110 nonhomologous proteins used to build segmental dictionaries contain the melittin and APPI, but I believe that the probability of generating a native-like melittin structure or an APPI structure by randomly assembling structural fragments from the segmental dictionaries is very small. For instance, melittin has 26 residues, and thus has 26 possible random conformational mutation sites. The number of possible conformation combinations for a dipeptide random search can be computed from the size of the dipeptide segmental dictionary. The lower bound of this number is about $(24{,}000/400)^{26} \approx 60^{26} \approx 10^{47}$; here, 24,000 is the number of dipeptide conformations in the dictionary. Notice that I have not counted the random perturbation at each (ϕ, ψ) yet in this estimation. If the random perturbation at each (ϕ, ψ) in Equation 16 is properly counted, the estimated number of possible conformational combinations has to be multiplied by an additional astronomical factor. Therefore, it is unlikely to generate randomly a native-like structure from the segmental conformation dictionaries in a feasible period of computation without the genetic algorithm optimization. This is because all its parts have to have correct conformations in order for a structure to be native-like.

Visualizing the conformational population as a subensemble of all the accessible conformations of a protein, the conformational species in this subensemble span across the entire conformational space of the system. The genetic optimization process describes a dynamic evolution process of this subensemble, in which the collective correlation among the conformation species governs the behavior of migration of the subensemble in conformational space, like a swarm of bees heading to a new hive, until this evolving subensemble is attracted in the basin of the energy landscape of the system. I would like to point out that despite the potential usefulness of the genetic algorithm conformational search method, the folding process computed by the current genetic algorithm RRM may have nothing to do with the real folding process. The real protein folding is a spontaneous dynamic process that is governed by all kinds of interactions between a protein and its surrounding environment. It is a process that involves the free energy gradients rather than the values of the free energy of the system. On the other hand, if the thermodynamic hypothesis is valid, the final folded conformation of a protein will not depend on the path along which it folds.

Acknowledgments

I have benefited greatly from the comments by reviewers, one of whom kindly corrected language errors in the original manuscript and brought the paper by Pincus et al. (1982) to my attention. Dr. Robert Spangler kindly polished the writing of the manuscript. I thank Dr. R. Parthasarthy and Dr. G. Snyder for their encouragement and discussions. The Pittsburgh Supercomputing Center and the computer center of State University of New York at Buffalo (Dr. F. Rens) have generously provided the necessary computer resources. My research has been supported in part by the Department of Biophysics of SUNY at Buffalo. Partial support from NIH (grant GM26715) through Dr. Snyder is also acknowledged.

References

Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., & Weng, J. (1978). Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Application* (Allen, F.H. & Sievers, R., Eds.), pp. 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science 181*, 223–230.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Jr., Rodgers, J.R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977). Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol. 112*, 535–542.

Blommers, M.J.J., Lucasius, C.B., Kateman, G., & Kaptein, R. (1992). Conformational analysis of a dinucleotide photodimer with the aid of the genetic algorithm. *Biopolymers 32*, 45–52.

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

Brooks, C.L., III, Karplus, M., & Pettitt, B.M. (1988). *Protein: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics. Advances in Chemical Physics*, Vol. LXXI. John Wiley and Sons, New York.

Bryngelson, J.D. & Wolynes, P.G. (1990). A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers 30*, 177–188.

Chan, H.S. & Dill, K.A. (1989a). Compact polymers. *Macromolecules 22*, 4559–4573.

Chan, H.S. & Dill, K.A. (1989b). Effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys. 92*, 3118–3135.

Chan, H.S. & Dill, K.A. (1990). Origins of structure in globular proteins. *Proc. Natl. Acad. Sci. USA 87*, 6388–6392.

Chan, H.S. & Dill, K.A. (1991). Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem. 20*, 447–490.

Chan, H.S. & Dill, K.A. (1993). The protein folding problem. *Phys. Today Feb.*, 24–32.

Covell, D.G. & Jernigan, R.L. (1990). Conformations of folded proteins in restricted spaces. *Biochemistry 29*, 3287–3294.

Crippen, G.M. & Viswanadhan, V.N. (1984). A potential function for conformational analysis of proteins. *Int. J. Pept. Protein Res. 24*, 279–296.

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

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

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

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, New York.

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

Heringa, J. & Argos, P. (1991). Side-chain clusters in protein structures and their role in protein folding. *J. Mol. Biol. 220*, 151–171.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor.

Huyghues-Despointes, B.M.P. & Nelson, J.W. (1992). Stabilities of disulfide bond intermediates in the folding of apamin. *Biochemistry 31*, 1476–1483.

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

Lau, K.F. & Dill, K.A. (1989). A lattice statistical mechanics model of the conformational and sequence space of proteins. *Macromolecules 22*, 3986–3997.

Lau, K.F. & Dill, K.A. (1990). Theory for protein mutability and biogenesis. *Proc. Natl. Acad. Sci. USA 87*, 638–642.

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

Levitt, M. & Warshel, A. (1975). Computer simulation of protein folding. *Nature 253*, 694–698.

McCammon, J.A. & Harvey, S.C. (1987). *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Oobatake, M. & Crippen, G.M. (1981). Residue-residue potential function for conformational analysis of proteins. *J. Phys. Chem. 81*, 1187–1197.

Pincus, M.R., Kalusner, R.D., & Scheraga, H.A. (1982). Calculation of the three-dimensional structure of the membrane-bound portion of melittin from its amino acid sequence. *Proc. Natl. Acad. Sci. USA 79*, 5107–5110.

Ramachandran, G.N. & Sasisekhharan, V. (1968). Conformation of polypeptides and proteins. *Adv. Protein Chem. 23*, 283–437.

Shakhnovich, E.I. & Gutin, A.M. (1990). Implication of thermodynamics of protein folding for evolution of primary sequences. *Nature 346*, 773–775.

Sippl, M.J., Hendlich, M., & Lackner, P. (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 Sci. 1*, 625–640.

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

Srinivasan, N., Sowdhamini, R., Ramakrishnan, C., & Balaram, P. (1990). Conformations of disulfide bridges in proteins. *Int. J. Pept. Protein Res. 36*, 147–155.

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

Taketomi, H., Kano, F., & Go, N. (1988). The effect of amino acid substitution on protein-folding and -unfolding transition studied by computer simulation. *Biopolymers 27*, 527–559.

Taketomi, H., Ueda, Y., & Go, N. (1975). Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pept. Protein Res. 7*, 445–449.

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

Terwilliger, T.C. & Eisenberg, D. (1982). Structure of melittin. *J. Biol. Chem. 257*, 6010–6022.

Tuffrey, P., Etchebest, S., Hazout, S., & Levery, R. (1991). A new approach to the rapid determination of protein side chain conformations. *J. Biomol. Struct. Dyn. 8*, 1267–1289.

Unger, R., Harel, D., Wherland, S., & Sussman, J.L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins Struct. Funct. Genet. 5*, 355–373.

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

Wilson, C. & Doniach, S. (1989). A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins Struct. Funct. Genet. 6*, 193–209.