PROTEINS: Structure, Function, and Genetics 6:193–209 (1989)

A Computer Model to Dynamically Simulate Protein Folding: Studies with Crambin

Charles Wilson[1] and Sebastian Doniach[2]

[1]Graduate Group in Biophysics, Department of Biochemistry and Biophysics, University of California, San Francisco, California 94143 and [2]Department of Applied Physics, Stanford University, Stanford, California 94305

ABSTRACT **The current work describes a simplified representation of protein structure with uses in the simulation of protein folding. The model assumes that a protein can be represented by a freely rotating rigid chain with a single atom approximating the effect of each side chain. Potentials describing the attraction or repulsion between different types of amino acids are determined directly from the distribution of amino acids in the database of known protein structures. The optimization technique of simulated annealing has been used to dynamically sample the conformations available to this simple model, allowing the protein to evolve from an extended, random coil into a compact globular structure. Many characteristics expected of true proteins, such as the sequence-dependent formation of secondary structure, the partitioning of hydrophobic residues, and specific disulfide pairing, are reproduced by the simulation, suggesting the model may accurately simulate the folding process.**

Key words: protein folding, simulated annealing, empirical potentials, Monte Carlo dynamics

INTRODUCTION

Despite years of both experimental and theoretical study, predicting the three-dimensional structure of a protein from its primary sequence is still a major unsolved problem. The need for a solution has increased in recent years with advances in DNA sequencing technology: the rate at which protein sequences are being determined by molecular biologists far exceeds the rate at which crystallographers can solve protein structures. While empirical force fields have been developed which can accurately predict the conformation of small molecules or subtle changes in folded protein structures (via molecular mechanics or molecular dynamics), these descriptions have not been useful for the simulation of protein folding because their highly dimensioned potential surfaces are characterized by many local minima. Solving the protein folding problem with these force fields would require the global optimization of thousands of variables which are nonlinearly coupled to each other. Early computer studies by Levitt and Warshell,[1] Tanaka and Scheraga,[2] and Kuntz et al.[3] showed the advantages of using simplified protein models. By limiting the number of structural degrees of freedom to one or two per amino acid, the conformation space which must be sampled to arrive at the final folded structure is significantly smaller than that available to a full-atom model of the protein.

Over the last 10 years, little new work has been done to explore the potential power of simplified models like those described in these initial studies. There are several reasons why this basic approach should be reconsidered. (1) The size of the protein structure data base which is used to develop and test these models has increased significantly. (2) Several new approaches to multivariate optimization problems, such as simulated annealing and neural networks, have been developed which may solve the local minimum problem for simplified models. (3) The speed of computers available for these studies has increased by several orders of magnitude, allowing a more complete search of conformation space (their speed would still have to improve by a factor of $>10^{10}$ if molecular dynamics on a full-atom model were used to fold a protein!). Historically, these models fell out of favor after it was shown that unreasonable, built-in biases in the early models allowed them to work for specific test proteins.[4] Hagler and Honig[4] demonstrated that an extremely simplified model for bovine pancreatic trypsin inhibitor (BPTI), in which the protein consisted of only glycine and alanine residues, produced a folded structure which was as correct as that generated by the Levitt–Warshel model. In addition, they showed that many topologically important features of BPTI, such as the 180° twist in its β-sheet, were not reproduced by any of the computer folding attempts. They conclude that by assuming certain properties about the starting conformation and manipulating terms in the force field used to evaluate different conformations, it is not hard to produce folded struc-

Received April 21, 1989; accepted June 13, 1989.

Address reprint requests to Charles Wilson, Graduate Group in Biophysics, Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94143-0448.

tures that bear some resemblence to a test protein. An important unanswered question, however, is whether a well-conceived, unbiased simplified model can contain enough structural and energetic information to truly predict protein structure.

The current work describes a simple representation for the structure and energetics of a protein. We have tried to avoid the potential problems of unwarrented bias in the model by relying completely on the statistical preferences observed in the data base of known crystal structures to derive all aspects of the model. Completely random starting structures are minimized using simulated annealing. In no way is the crystal structure of our test protein (crambin) used to define the model. Hopefully, therefore, our findings can be generalized to a large extent to other proteins. Several results indicate that the model is able to reproduce the characteristics of a folded protein in a sequence-dependent way.

Crambin was chosen as a test protein because its crystal structure is known to high resolution and it is one of the smallest proteins with significant amounts of secondary structure. Different aspects of the correct crambin structure can be analyzed and compared to the results of the folding simulation. The agreement or disagreement between the two sets of data can then be used to assess the quality of the model. In a number of ways, listed below, the folding simulation is able to reproduce the features of the true crambin structure. (1) α-Helices and β-strands develop where they occur naturally in the crystal structure. (2) The structures produced by the minimization are reasonably similar to the true structure in terms of their rms deviation or distance matrix error (to our knowledge, their average deviation is the lowest ever obtained by the simulated folding of a protein). (3) The formation of certain disulfides, corresponding to the true disulfides, is significantly favored over nonnative disulfides. The simulation with the true crambin amino acid sequence and a variety of control experiments using related sequences suggests that these results are not dependent on fortuitous biases within the model. For instance, completely different structures result when the crambin sequence is randomized, maintaining the same amino acid composition. The tendency for secondary structure to form is largely lost and that which does form is quite different when folding individual segments of the sequence, as might have been expected from studies with small peptides. Starting from different random models, simulated annealing produces a collection of closely related folded structures, indicating the ability of the method to avoid many local minima during the journey down the potential surface toward the global minimum. Several properties of the native folded models, such as the tendency for hydrophobic residues to partition into the interior of the protein, indicate that the folding process produces "protein-like" structures. Taken together, these results strongly encourage the use of simplified models for protein structure, both for the purpose of protein structure prediction and as a means of studying the energetics and dynamics of protein folding.

While this model certainly does not solve the protein folding problem, it suggests that a large part of the physics which is important for defining protein structure can be modeled by united atom residues and simple pairwise potentials.

METHODS

Several models have been used in the simulation of protein folding. Lattice models, in which each residue is constrained to lie at a lattice point with appropriate connections to adjacent residues, have been used because it is computationally possible to generate a significant fraction of the allowed conformations. The energy of a given conformation for a lattice model is generally calculated as the sum of interactions between adjacent or near-neighbor lattice sites, the value of each interaction determined by the type of interacting residues. Go and Taketomi[5] represented space as a two-dimensional lattice and were able to hypothesize a role for short- and long-range forces in the folding transition. The highly simplified geometry of such models, however, lacks the basic structural properties of true proteins and thus compromises the accuracy of the model for the simplicity with which conformations may be generated.

The models used for simulations in the present work have attempted to maintain a minimum of structure while preserving the basic geometry of the polymer. All backbone bond lengths and bond angles were held at ideal values[6] and the dihedral angle for the peptide bond was fixed in the trans conformation (180°). The side chains for residues were represented as a point whose position was determined by averaging the side chain centroids of the residues in 25 proteins in the Brookhaven Protein Data Bank (PDB). The only variables determining the conformation of each residue are thus the dihedral angles (phi and psi) defined by the bonds on either side of the α-carbon. For the current work, approximately 100 proteins in the PDB were used to generate probability plots of allowed phi–psi pairs (see example in Fig. 1). Contours in the observed phi–psi distributions were then used directly to produce a set of randomly generated phi–psi pairs for each residue type. This set of phi–psi values was subsequently used to assign a conformation at random for each residue. With the exception of proline and glycine, the only significant differences between the various residues were the relative populations of the α-helix and β-sheet regions of their phi–psi probability plots (Table I).

Empirical potentials describing the interaction between each amino acid pair were derived directly

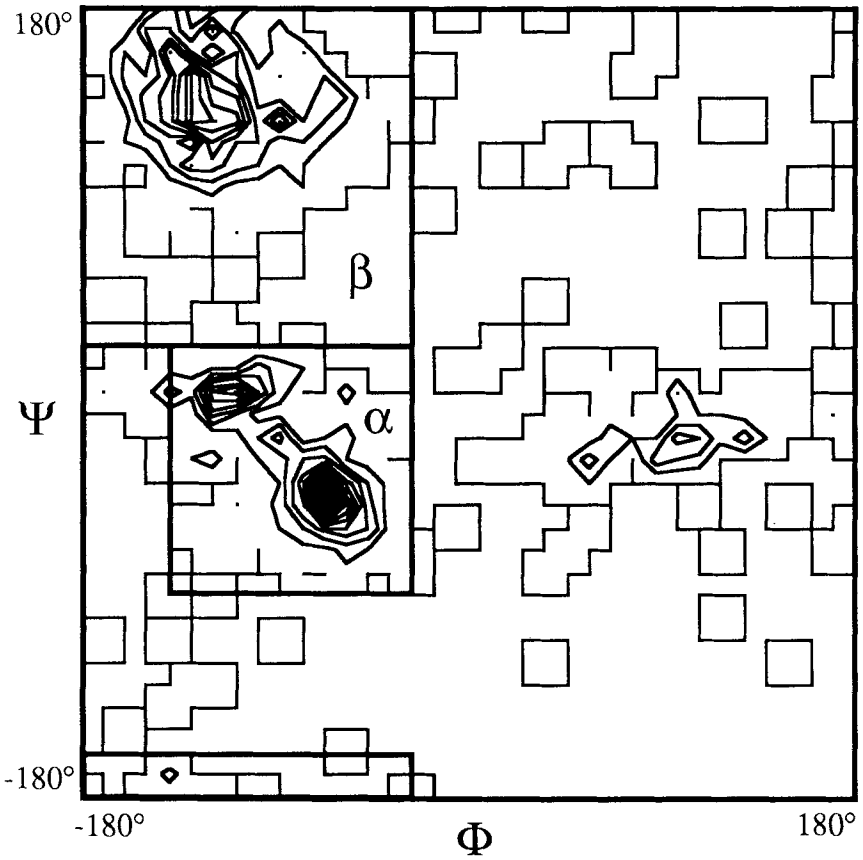

Fig. 1. Phi-psi probability plot for valine. Phi–psi probability plots were calculated for each amino acid as described in Methods. The probability of a residue having a set of phi–psi values was calculated for each 10° × 10° pixel of phi–psi space. Probability contours are drawn for the valine plot in intervals of 0.0015 (twice the average value for all pixels). The regions defined as α and β conformations are enclosed and labeled.

from the distribution of residues observed in known protein structures. Tanaka and Scheraga[2] and Crippen and Viswanadhan[7] have developed potentials with a similar basis in somewhat different ways. In the current work, the C_α–C_α distances of all residue pairs in 100 PDB proteins were tabulated and used to generate histograms of the number of occurrences versus distance for each amino acid pair. The occurrences in the 15–16 Å range were used as a reference level since no specific contacts were assumed to occur at this large separation. The observed distributions of residue–residue distances were converted into free energies, assuming a Boltzmann distribution.

Potentials calculated in this way should include information about both long- and short-range packing interactions. The tendency of a residue to be buried can be inferred by examining the pair potentials for that residue with all other residues. Buried, hydrophobic residues consistently form many short-range contacts and show favorable interaction energies with a large number of other residues. Exposed residues on average form fewer close contacts (and thus have less favorable short-range interaction energies), although certain pairs may interact quite favorably. These tendencies can be demonstrated by comparing the pairwise interactions of leucine and lysine with other amino acids (Fig. 2). Leucine, a commonly buried hydrophobic residue, interacts favorably with most amino acids at 4–10 Å. Lysine, on the other hand, shows much weaker interactions on average, despite strong tendencies to contact aspartate and glutamate residues. The potentials are affected by packing in addition to the general hydrophobic partitioning of amino acids. Since secondary structures generally associate in well-defined ways, they alter the potentials by favoring certain pairwise distances between residues. For example, two α-helices often align such that many of the distances from a residue in one helix to a residue in the other helix fall in the range of 8–10 Å. Many of the contacts made by β-strand residues are to residues on adjacent β-strands and these are typically 4–6 Å long. Specific short-range interactions, such as electrostatic attraction or repulsion between charged side chains, also perturb the potentials in a se-

Table I. Phi–Psi Probabilities*

	Distribution					
	α		β		Neither	
Amino acid	#	%	#	%	#	%
Alanine	567	54.5	312	45.0	162	15.6
Arginine	174	48.5	130	36.2	55	15.3
Asparagine	224	41.1	168	30.8	153	28.1
Aspartate	324	45.2	214	29.8	179	25.0
Cysteine	111	39.1	121	42.6	52	18.3
Glutamine	198	45.8	162	37.5	72	16.7
Glutamate	353	58.8	129	21.5	118	19.7
Glycine	242	21.5	183	16.3	701	62.3
Histidine	103	40.2	85	33.2	68	26.6
Isoleucine	234	39.1	268	44.7	97	16.2
Leucine	420	47.4	318	35.9	148	16.7
Lysine	420	51.7	205	25.2	188	23.1
Methionine	88	47.8	63	34.2	33	17.9
Phenylalanine	187	41.2	183	40.3	84	18.5
Proline	178	35.6	244	48.8	78	15.6
Serine	414	42.9	377	39.0	175	18.1
Threonine	299	39.7	333	44.2	121	16.1
Tryptophan	78	43.1	73	40.3	30	16.6
Tyrosine	148	35.1	211	50.0	63	14.9
Valine	316	35.7	415	46.8	155	17.5

*Phi–psi probability plots were calculated for each amino acid type as described in Methods. Phi–psi space was broken into three regions termed α ($-140°<\Phi<-30°, -80°<\Psi<30°$), β ($-180°<\Phi<-30°, \Psi<-160°$ or $\Psi>30°$), or neither (all remaining areas). These labels are for reference only and do not indicate that all phi–psi pairs within each region were obtained for a particular secondary structure (i.e., α-helix or β-sheet). The table lists the number of occurrences and fraction of residues found in each region of phi–psi space for each amino acid type.

quence-dependent way. The complete set of potentials is available as supplementary material or as part of the folding program. A thorough analysis of the potentials should reveal many of the patterns that have been found by previous surveys of the protein structure data base. Our work, however, simply uses the potentials as an unbiased representation of the packing tendencies without attempting to explain them.

Since the potentials are based upon α carbon distances, they can include lower resolution structures from the PDB than those calculated by the method of Tanaka and Scheraga in which the position of all atoms must be well defined to determine if an interatomic contact has formed. To avoid the effect of near-neighbors along the sequence altering the distance potential without adding significant information about the conformation (since near-neighbors are covalently constrained to lie within a certain distance of each other), only the distances between residues separated by at least four other residues were included in the histograms. To calculate the energy of a given conformation, the distance between the ith residue and all others, $j = 1$ to $i-5$ and $i+5$ to N, were calculated and the corresponding energies were totalled.

The interaction between the side chains was initially ignored but early studies showed that this simplification was unreasonable. Using only α-carbon-based potentials, the model peptide tended to pack into a globule much denser than the native protein. The space that should be occupied by side chain atoms disappeared as the backbone potential was optimized. The pairing of two β-sheets brings many backbone atoms in close contact with each other while maintaining good separation between the side chains. Without explicit side chains, the initial model brought many residues into this potential well without allowing the subsequent addition of side chains. As such, the model was modified to include potentials based on side chain centroid-to-centroid distances. These potentials were calculated and applied as described above for the backbone potentials. The addition of these potentials significantly raised the average radius of gyration for the model protein, bringing it closer to that observed for the true structure.

A representative set of potentials is shown in Fig. 2. For this and other potentials, the distribution appears to be the sum of two wells, one centered at 4–6 Å (corresponding to residues in β-conformation) and another at 8–10 Å (α-helical residues). Since conformation appeared to significantly affect the proper choice of a potential, the model was expanded to include separate potentials for each residue pair and residue conformation (based on phi–psi value) to determine the energy of a given structure. A single potential independent of residue type was applied to residue pairs separated by four residues (i–$i+4$ and i–$i-4$ interactions). This potential was determined using the procedure described above (a histogram of the distance between all i–$i+4$ residue pairs, segre-

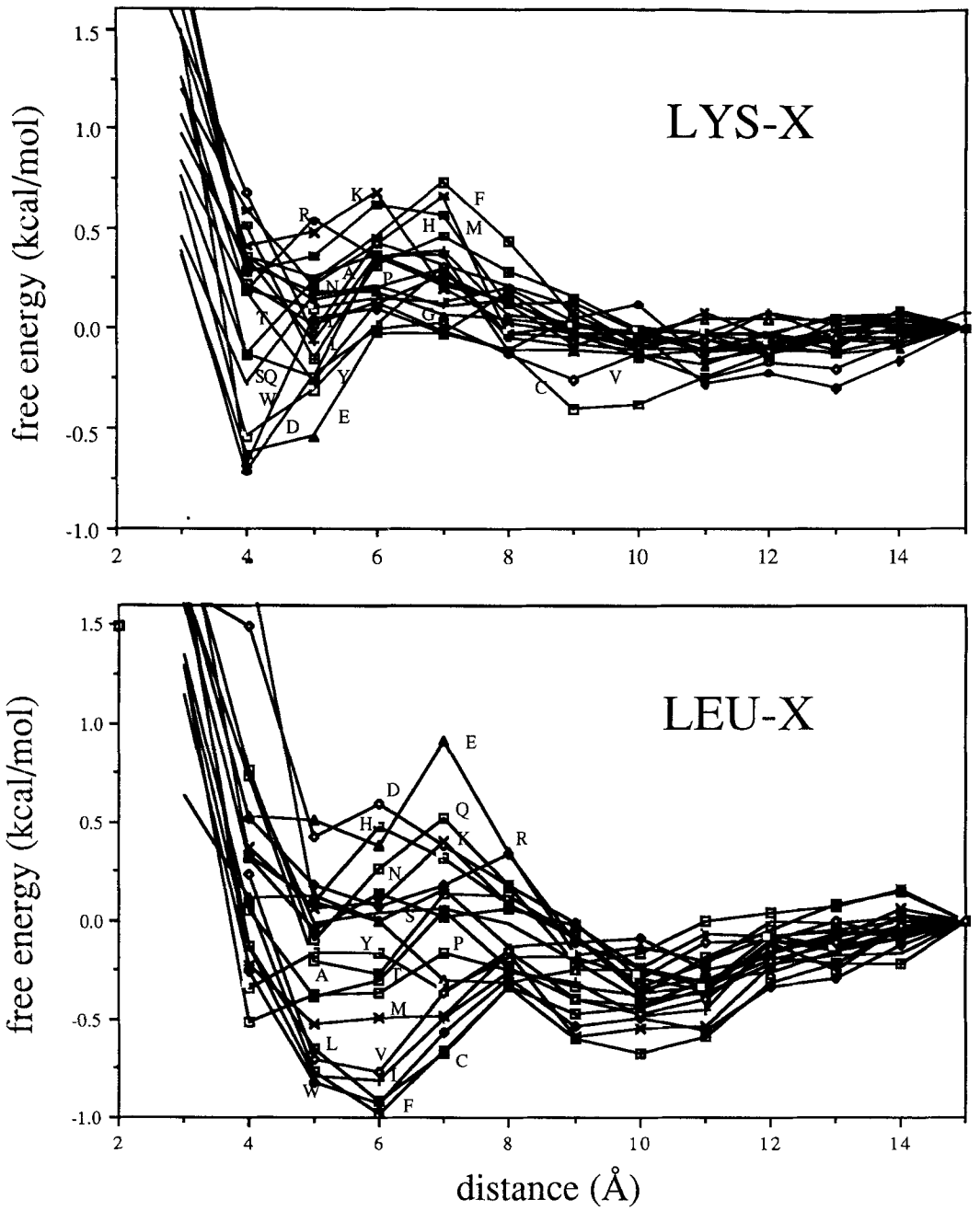

Fig. 2. Potential for interaction with leucine and lysine. Upper panel: the side chain distances from every lysine to every other amino acid side chain were calculated for 100 PDB structures and used to produce a distance-dependent potential for the Lys-X interaction as described in Methods. Each potential is labeled by the corresponding one-letter code for the amino acid. Negatively charged and other polar amino acids (D,E,W,S,Y) are strongly attracted to lysine, while positively charged amino acids (R,H,K) and some hydrophobic amino acids (F,M) are repelled. Lower panel: As above, but for the Leu-X interactions. As expected, the hydrophobic amino acids (F,W,C,I,L,V,M) interact favorably with leucine while the charged amino acids (E,D,H,K,R) interact with it poorly.

gated by conformation, was generated and converted into a free energy function of distance). There are approximately 12,000 residues in the data base used for this study and thus close to 12,000 i–$i+4$ distances to be measured. If separate potentials for each amino acid pair were used, there would on average be less than 20 data points in the histogram for each potential. We therefore combined all i–$i+4$ distances to produce a single potential which was used for all amino acid pairs.

Given a way to calculate the energy for any protein conformation, we required a means for adjust-

ing the backbone angles to minimize the calculated energy. Simulated annealing is an optimization procedure by which high-energy or highly disordered systems may be "cooled" in a controlled way that efficiently produces a system whose energy is close to that of the global minimum.[8] It has recently been used as a method for generating protein structures which satisfy specific interproton distance constraints.[9] Its ability to rapidly sample over a highly convoluted energy surface makes it an attractive candidate as a protein conformation searching technique. As described by Kirkpatrick et al.,[8] simulated annealing can be carried out as a Monte Carlo dynamics procedure in which the "temperature" of the system monotonically decreases during the course of the simulation.[10] In a run, the temperature is initially high such that the conformation of the system is rapidly changing and sampling a large volume of conformation space. As the run proceeds, the temperature is lowered slowly to allow the system to fall into a basin of attraction without becoming locked into the first local minimum reached. The specific heat of the system provides a good measure of the conformational transitions that are occurring and can be used to limit the rate of cooling to prevent premature freezing into local minima. The model described above was incorporated into a Monte Carlo dynamics program, written in FORTRAN, which was compiled and run on an FPS 264 (available in VAX and FPS versions, together with all the analysis programs used, upon request from CW). Simulated annealing was carried out using a schedule for cooling similar to that suggested by Kirkpatrick et al.[8] for the simulation of computer chip design.[10] During the run, a residue and a new phi–psi pair to replace the current value were chosen at random and the new conformation and corresponding energy were calculated. A thermal equilibrium was attempted at each temperature by applying the basic Metropolis algorithm.[10] In the Kirkpatrick annealing schedule, the relative temperature, kT, is lowered by a constant factor if either (1) 10 moves per site have been accepted since the last temperature change, or (2) 100 moves per site have been attempted without the required number of accepted moves. The first condition will be satisfied early during the run and allows the system to start at an arbitrarily high temperature and to rapidly cool down to a temperature at which true minimization can begin. Later, the second condition is satisfied and is used to signal the end of the run. The system is considered frozen and the run ended when the second condition is applied consecutively 3–10 times. For the current studies with crambin (46 residues), each annealled structure required approximately 5–30 minutes of CPU time on the FPS 264 (corresponding to approximately 1–6 minutes on the Cray X-MP and 1–6 hours on a VAX-780).

Knowledge of the secondary structure may be easily incorporated into the model by constraining the choice of dihedral angles to values within the region of phi–psi space corresponding to either α-helix or to β-sheet. For simulations in which the secondary structure was considered known, the phi–psi values for α-helical residues were constrained to ($-140°<\Phi<-30°, -80°<\Psi<30°$) and those for β-sheet residues were limited to ($-180°<\Phi<-30°$, $30°<\Psi<-160°$). (The probability of occurrence of phi–psi pairs was obtained from the observed phi–psi probability plots for each amino acid type.) Since this constraint limits the possible effects of altering the ordered residues (only a relatively small change in conformation is allowed), the program was written such that selection of these residues during the Monte Carlo procedure was slightly disfavored, typically chosen 60% as often as unassigned residues.

The structure of crambin has been solved to 0.945 Å resolution with an R-value of 0.129, one of the most accurate protein structures ever determined.[11,12] The native protein contains a pair of α-helices, labeled here as A_1 (residues 7–19) and A_2 (residues 23–29), covalently joined to each other by a disulfide between residues 16 and 26. At either end of this helix pair are two short strands of β-sheet, B_1 (residues 1–4) and B_2 (residues 32–35), joined by a disulfide between residues 4 and 32 in an antiparallel sheet. The C-terminus (residues 36–46) has no regular structure, although a reverse turn is defined by residues 41–44. [Secondary structure assignments were taken from the Protein Data Bank entry of the crambin coordinates [1CRN).] Cys-40 is covalently linked to Cys-3, thus bringing the terminus to lie adjacent to the first β-sheet. In simulations in which the secondary structure was considered known, the above assignments for helix and sheet were used.

RESULTS AND DISCUSSION

Does Simulated Annealing Solve the Local Minimum Problem for a Simplified Protein Model?

Simulations using the simplified model described above were carried out to test the effectiveness of simulated annealing as a conformation searching technique. Fixed-temperature Monte Carlo simulations were done at several temperatures, starting from random conformations. The results were compared to a simulated annealing minimization in which the temperature was lowered periodically by 5% according to the schedule described above. Twenty structures were produced during each run and average characteristics for all simulations are shown in Table II. The structures produced by the simulated annealing run had the lowest average energy and radius of gyration of structures from any of the simulations. The model peptide tended to fall into a compact structure while simultaneously avoiding the formation of bad overlaps. The contact

Table II. Characteristics of Fixed Temperature and Annealed Simulations*

Temperature	Energy (kT)	Radius of gyration (Å)	Total contacts	rms native (Å)	DME native (Å)	DME internal (Å)	DME average (Å)
0.5	2531.6	9.68	781.2	8.15	5.34	5.03	4.06
1.0	2507.2	9.90	755.6	7.95	5.54	5.03	4.30
2.0	2528.5	9.65	784.6	8.01	5.36	4.94	4.14
5.0	2529.8	9.77	754.9	7.72	5.16	4.95	3.86
10.0	2509.7	9.99	727.0	8.36	5.51	5.35	4.08
20.0	2690.8	12.24	590.5	8.34	7.01	7.04	5.29
100.0	3357.3	17.47	469.7	12.42	14.28	7.91	13.46
Simulated annealing	2349.7	9.15	828.1	7.54	4.76	4.21	3.76
Crystal structure	3192.9	9.70	876.0	—	—	—	—

*Monte Carlo dynamics was run at the indicated fixed temperatures. For the simulated annealing run, the temperature was initially set to 100, and gradually lowered. rms native refers to the average rms deviation of each model structure from the PDB crystal structure. DME indicates the average distance matrix errors between each model structure and the crystal structure (native) and between each model structure and every other structure (internal). The average DME is the distance matrix error calculated by comparing the contact map of the crystal structure to the average contact map generated from all the model structures of a given simulation.

maps for the structures can be averaged and those for some of the simulations are shown in Figure 3. For the nonannealed structures produced at low temperatures, contacts formed early between the residues at either terminus and the remainder of the peptide. These initial folds were sufficiently stabilizing that they could not be broken to allow more compact and lower energy conformations to be made. At higher temperatures, these nonannealed structures failed to form many long-range contacts and most contacts occurred between residues separated by less than six residues. The annealing simulation contact map, however, showed that extensive contacts could be formed between pairs of helices and sheets to produce a low-energy structure (Fig. 4).

If the potentials used to calculate the energy of each conformation are an accurate approximation of the energetics for a true protein, the energy calculated for the native structure should be a global minimum, lower than all other possible conformations. For all models attempted, however, several structures were found with calculated energies lower than that for the native. There are several possible explanations for this problem. A major simplifying assumption of our model is that the effect of solvent can be treated implicitly using potentials between pairs of amino acids, rather than by explicit interactions between individual residues and water molecules. Since most potentials are favorable in the 5–15 Å range, the structures tend to be uniformly compressed to maximize the number of amino acid contacts that fall into this distance range. The average radius of gyration for all models was approximately 0.5 Å smaller than that for the native structure. In several of the structures produced by annealing, the characteristic "L"-shape of the native crambin (formed on one face by the helix pair and on the other by the β-sheet) appears to be compressed into a "V"-shape. This compression adds several helix–sheet interactions to the native helix–helix and sheet–sheet contacts, thus lowering the calculated energy relative to the native conformation. In reality, specific interactions of the solvent with the interface between the helix pair and the β-sheet may be more favorable than those formed when the protein is compressed. Unless solvent is included explicitly, however, there is no obvious way to prevent this compression without removing the driving force for folding. Other possible inaccuracies in the potentials that have been used will be discussed later.

There are two basic requirements for any simulated folding process: (1) random starting conformations should converge to a single minimized structure, and (2) the structure of the model at its global minimum should be identical to the native, experimentally observed structure. In practice, these goals are related and they have never been completely satisfied by a computer model. If random starting structures do not converge, it is unlikely that the global minimum is one of the structures in the final set. Without the global minimum defined by the model to test against the native structure, it is impossible to determine the accuracy of the model. To assess the ability of the annealing procedure to converge to a single minimum (self-convergence) and the similarity of the final set of structures to the native protein, the minimized structures were compared using two similarity metrics. The most common measure for comparing protein crystal structures, the rms deviation obtained by optimally translating and rotating one structure onto another (rms), has questionable usefulness when comparing two structures that are similar to low resolution.[13]

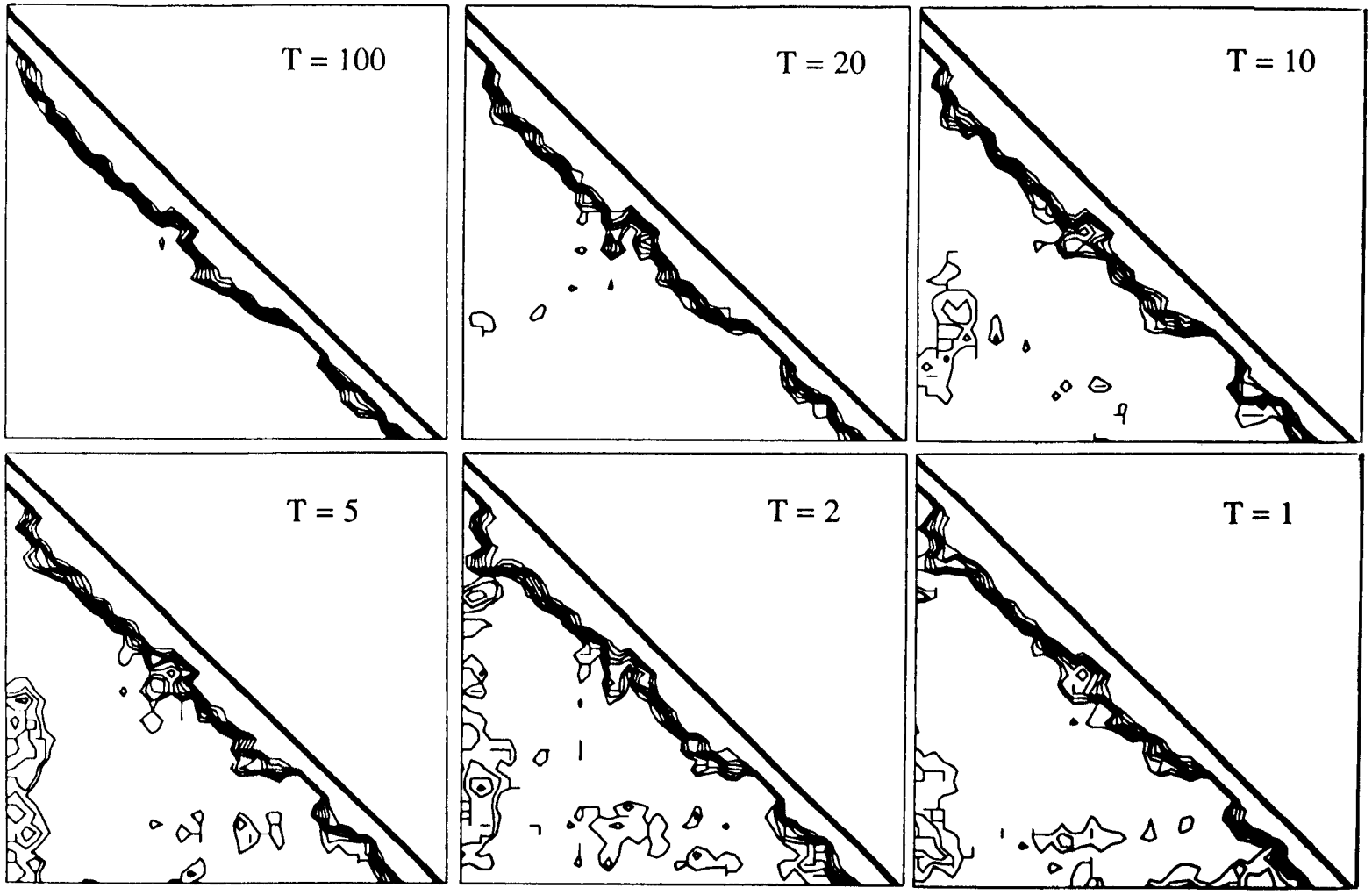

Fig. 3. Average contact maps for simulations as a function of temperature. Average contact maps were generated for structures from several of the fixed-temperature (nonannealed) simulations. In a contact map the protein sequence runs from the N- to C-terminus down the vertical axis and left-to-right along the horizontal axis. Contacts were defined for each pair of residues whose α-carbon atoms were separated by less than 10 Å. The contour levels in the plots indicate the frequency with which contacts between residues at the intersection of each point occurred among the structures produced by each simulation.

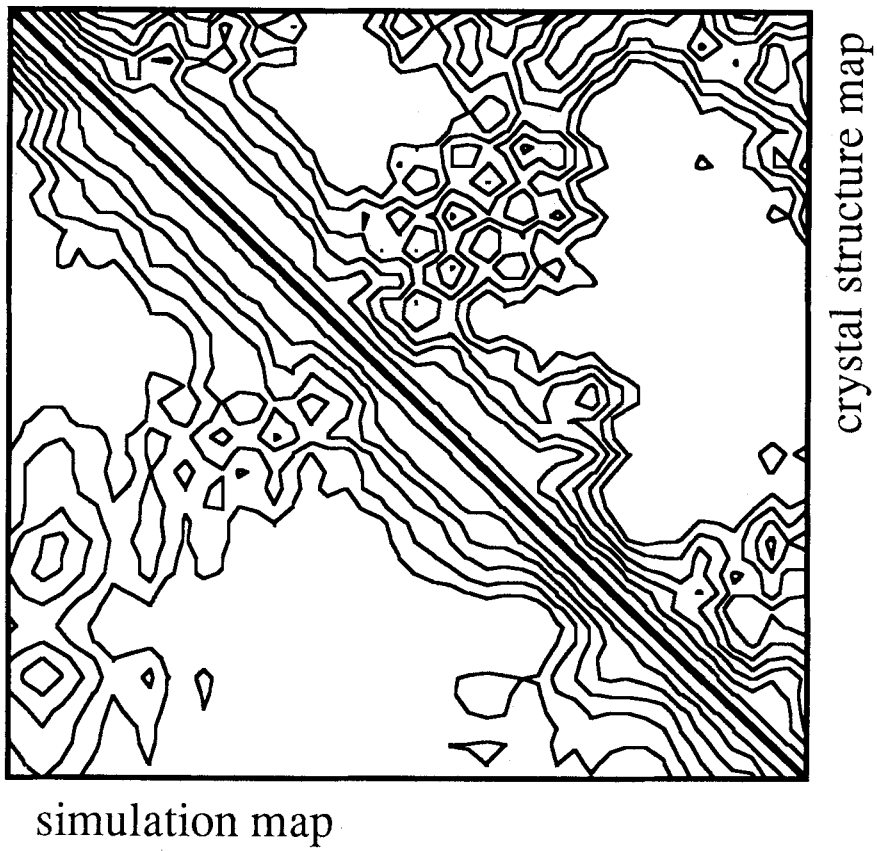

Fig. 4. Average contact map for the annealed model and the native structure. The average contact map calculated for the 100 structures produced by simulated annealing with explicit disulfides is shown in the lower triangle while the contact map for the native structure is shown in the upper triangle. The contours for both contact maps are based on the distance between α-carbons for each residue (with the shortest contacts enclosed by the highest contours).

Another common measure, the rms distance matrix error (DME), will determine the extent to which a pair of structures have similar patterns of contacts but ignores any differences in the chirality of the structures.[14] For instance, a left-handed and a right-handed helix may have a DME deviation approaching zero but a relatively high rms deviation. The DME deviation is typically less than the rms deviation by 30%. Both metrics were determined for the structures produced by the different minimization procedures (Table II). Figure 5 shows several sample model structures produced by simulated annealing, together with the crystal structure. One of these structures (fifth row) is especially interesting: while its general fold appears correct at first glance, the rms deviation is quite large (7.5 Å). The DME for this structure, on the other hand, is only 3.5 Å. Crambin is "L-shaped," with the paired α-helices and the β-sheet lying perpendicular to each other. The model structure has the same L-shape but is inverted. The DME metric, unlike the rms metric, is able to identify the general correctness of this fold (correct in the sense that the long-range contacts which energetically favor this conformation have formed).

To test the first criterion, self-convergence was estimated by calculating the average deviation between all pairs of structures from a given run. The deviation between structures produced by simulated annealing averaged 4.21 Å (DME), significantly less than that for all of the fixed temperature Monte Carlo simulations (4.94–7.91 Å). To estimate the ability to converge to the true global minimum, the deviation of the model structures from the crystal structure was calculated. The average deviation of the annealed structures from the native protein is 7.58 Å (rms), 4.76 Å (DME). [The model closest to the native structure had a deviation of 4.01 Å (rms), 3.17 Å (DME).] The deviation between any single structure and the crystal structure is thus only slightly more than the deviation between any pair of model structures. It is possible, therefore, that the inability to produce structures closer to the native is limited by the minimization procedure (a convergence problem), rather than by the potentials and the model. However, since the native state is not a

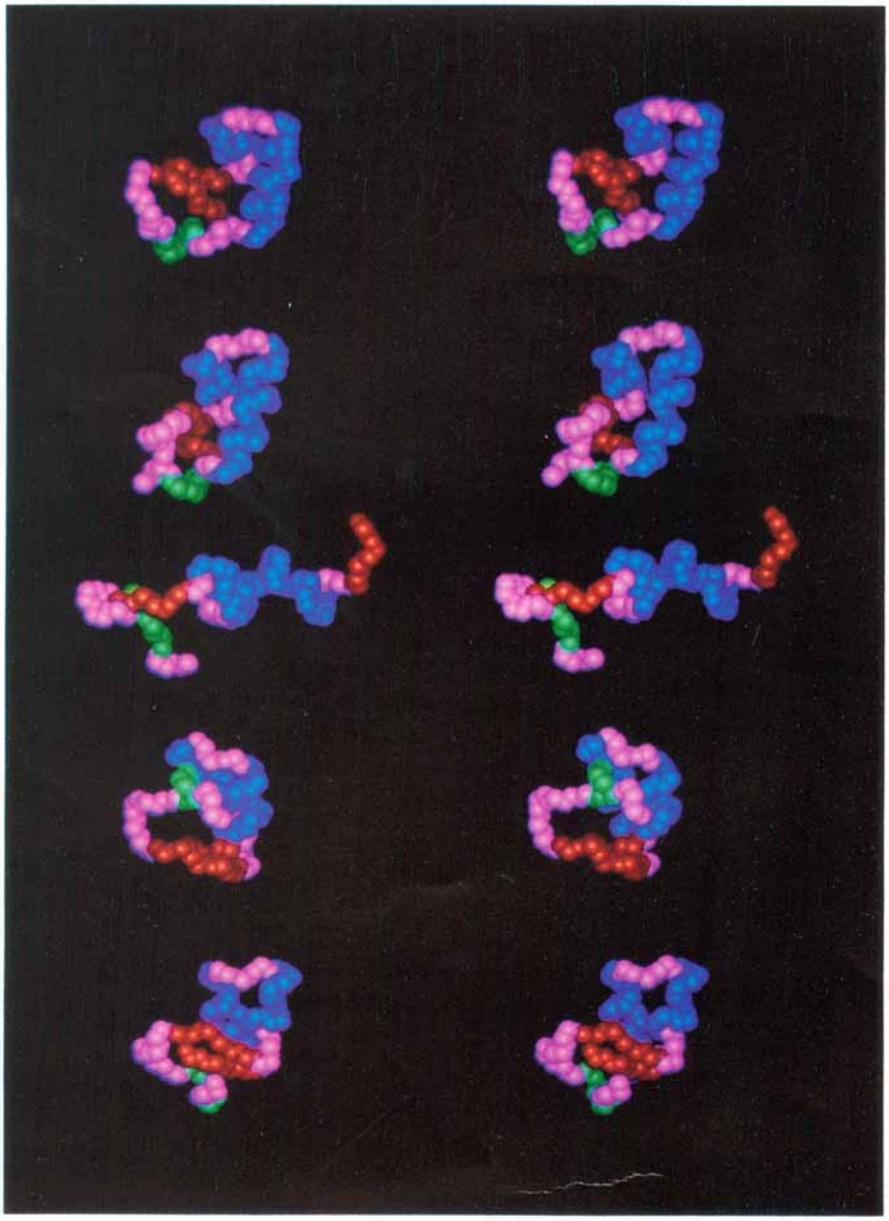

Fig. 5. Several minimized model structures and the native structure. A variety of model structures produced by minimization are shown in the same orientation as the crystal structure (α-helix = red, β-strand = blue, reverse turn = green). First row: The crystal structure (PDB name 1CRN). Second row: The "best" structure produced by minimization (rms = 4.01 Å, DME = 3.17 Å). The alignment of helices is nearly perfect (rms deviation <1.5 Å) and the overall construction of the model is correct, although the β-strands are poorly packed. Third row: The starting structure which yielded the best structure on minimization. Fourth row: The "worst" structure produced by minimization (rms = 9.59 Å, DME = 5.13 Å). The C-terminal 15 residues have packed against the open face of the helix pair rather than lying next to the β-sheet as they do in the native structure. Fifth row: An unusual structure produced by minimization. The rms deviation of the minimized structure from the native structure (7.5 Å) would not suggest that the predicted fold is similar to the true fold. The low DME deviation (3.5 Å), however, suggests that the model structure may have a pattern of contacts highly similar to the native. Inspecting the structures, it is clear that the overall shape and local contacts in the native structure have been reproduced by the model but that the two helices have been swapped.

global minimum as defined by the potentials, it will be impossible to obtain the correct structure by a complete search of conformation space. A contact map for the collection of minimized structures can be generated by averaging the individual contact maps for each structure. The distance matrix error of the average simulated structure with the native is 3.76 Å, significantly lower than the average error of any single structure with the native (4.76 Å). This suggests that while every structure differs in some way from the native, the average tends to fold in a way quite similar to the native. We are currently using distance geometry methods to rebuild a three-dimensional structure from the average contact map.

Does Minimization Produce "Protein-Like" Conformations?

The models produced by simulated annealing were analyzed in a variety of ways to determine if they are structurally similar to true proteins and to the crambin structure in particular. Hagler and Honig, in their critique of early protein-folding models, emphasize the need to examine the structural details of the model protein to determine the accuracy of the simulation. For the current work aspects of secondary structure, the hydrophobic effect, and disulfide bond formation were considered. As an additional test of the folding procedure, the crambin sequence was randomly shuffled to yield a sequence of identical composition but with no homology to crambin. Differences between minimized structures produced for the native and shuffled sequences can then be used to indicate the sequence dependence of the modelling.

Secondary Structure

Inspection of the phi–psi plots for amino acids shows that each have different biases toward α-helical or β-sheet conformation. These biases can be used in a simplistic way to predict secondary structure for a sequence. Assuming that the choice of phi–psi value for a residue is independent of all other neighboring residues, the probability of having a regular structure spanning a window of residues can be calculated as the product of the probabilities of forming that structure for each residue in the window. (A regular structure is defined here as a linear set of residues whose phi–psi values are all within the same region of phi–psi space.) The probability, P_i^a, that the ith resiude will lie in a type-a regular structure of length equal to or greater than w, may be calculated as

$$P_i^a = 1 - \prod_{n=i-w+1}^{2} (1 - \prod_{m=n}^{n+w-1} p_m^a)$$

where p_m^a is the probability that the mth residue will have a phi–psi value corresponding to type-a regular structure. (This procedure is similar to the Chou–Fasman secondary structure algorithm,[15] although the p_m^as which define the probabilities are determined quite differently.) This method was applied to the crambin sequence over a window of four residues and identified several regions with slight α-helix or β-sheet tendencies (Fig. 6). The tendencies calculated by this method are equivalent to those observed for the starting structures (the profiles obtained from nonmimimized structures were identical to those in Fig. 6). When simulated annealing was carried out with no assigned secondary structure, the observed regular structures occurred in the same regions as in the native structure and much more frequently than that predicted by the simple method described above. The tendency to form regular structures during the minimization seems to be initiated by a bias in the phi–psi values but enhanced during the simulated folding. Reasons for the enhancement of secondary structure are unclear. One possibility is that the formation of secondary structure allows residues to pack in defined ways and thus tends to increase the number of favorable contacts and lower the total energy. However, there was a wide range in the amount of regular structure formed in the minimized models (from 0 to 45% of all residues) and no correlation to the number of contacts or to the calculated energy could be found. Several models with no regular structure were produced with an equal number of contacts and energy as those models with large amounts of regular structure. Thus, while regular structures may be enhanced during folding, they cannot be said to be required for folding. A reasonable explanation for the enhancement of secondary structure is that the potentials favor certain pairwise distances (4–6 Å, 8–10 Å) and that one way (but not the only way) to produce folded structures with these distances is to form secondary structure (β-strand pairing creates residue–residue distances of 4–6 Å, α-helix pairing forms 8–10 Å contacts).

Prior to minimization, there are several regions that show tendencies toward both helix and sheet. After minimization, there is little overlap between the two types of structure, making it much easier to assign secondary structure to the sequence. A variety of prediction schemes have been developed that use local sequence information to assign secondary structure for each residue.[16,23] None of these techniques, however, is more than 70–80% accurate when tested against a collection of known structures. The inability of these methods to correctly predict secondary structure has been ascribed to long-range interactions which influence secondary structure formation but cannot be determined from a simple survey of the sequence.[17] By deleting segments of the sequence and attempting to fold a partial protein, one can estimate the importance of these long-range effects. Decapeptides correspond-

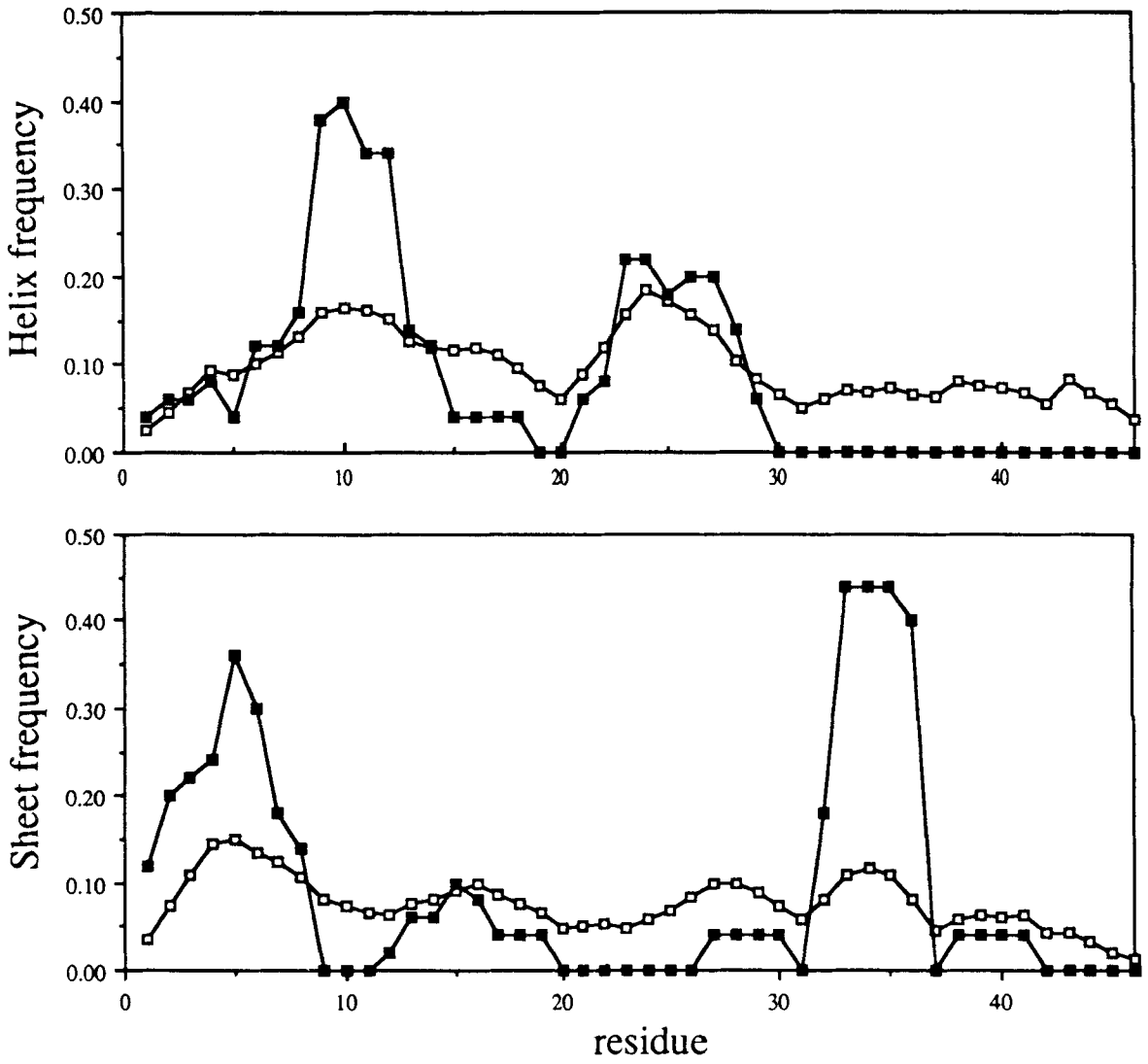

Fig. 6. Formation of secondary structure. The occurrence of regular structure, defined as four or more residues in either α-helical or β-sheet conformation, was tabulated for nonminimized starting structures (open squares), and for 50 structures produced by simulated annealing with no assigned secondary structure (solid squares). The formation of helix is limited to those regions in which significant bias for helix formation exists, corresponding, in fact, to the helical regions in the native structure (residues 7–19,23–29). The formation of β-sheet is significantly enhanced during minimization in regions overlapping the sequences of β-sheet in the native structure (residues 1–4, 32–35).

ing to regions of the crambin sequence were folded using the model and the resulting pattern of secondary structure was compared to that obtained when the entire protein was folded (Fig. 7). There are large differences between the secondary structure of a decapeptide in solution and that of a decapeptide in the context of the protein. These results demonstrate convincingly that the formation of secondary structure during the simulation requires information from a sequence as large as the protein itself. Attempting to model tertiary structure may take long-range interactions into account and thus improve the prediction of secondary structure.

The formation of the correct secondary structure would not be surprising if the starting conformations had this native structure built into them. The simplified models of both Levitt–Warshell and Hagler–Honig assume a starting conformation in which the BPTI peptide is completely extended. Since the folded BPTI structure consists mostly of extended β-strands, separated by turns, it is not surprising that the simulations do as well as they do. In contrast, our model makes no assumptions about the starting structure as phi–psi values for each residue are chosen at random from a phi–psi probability plot. Any starting structure forced onto the model is quickly lost because of the randomizing effect of high temperature in the initial stages of refinement.

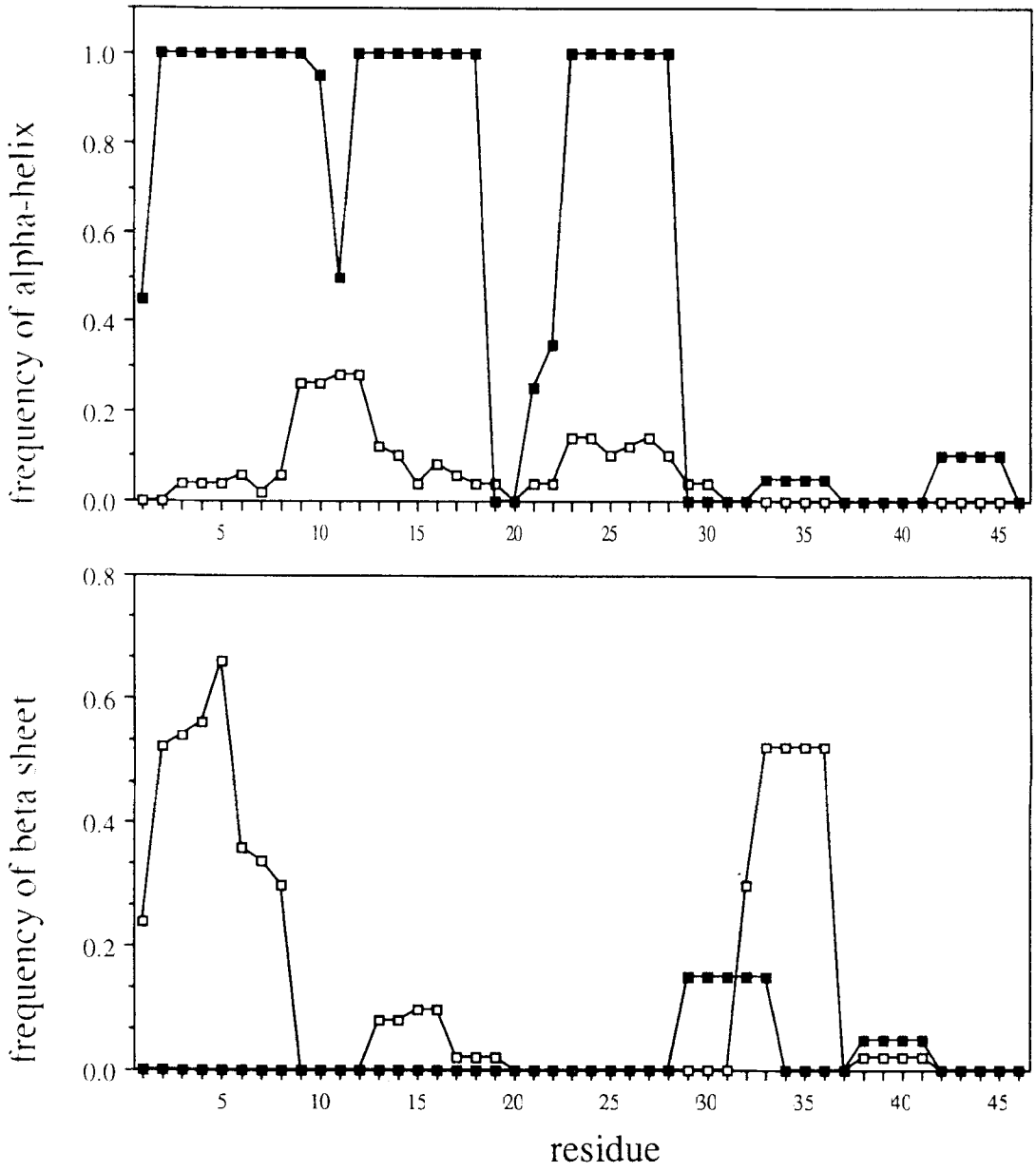

Fig. 7. Formation of secondary structure by peptide fragments. The effect of sequence context on the formation of secondary structure was determined by breaking the crambin sequence into a series of nonoverlapping decapeptides. Twenty structures were generated by minimizing each decapeptide as described in the text. The resulting secondary structure (solid squares) is compared to that calculated for the intact crambin minimization (open squares). The frequency of helix formation is significantly increased while β-strands are virtually eliminated.

This can be shown by comparing simulations in which either a completely α-helical structure or a completely extended (β-strand-like) structure is used as the starting point for refinement. No significant changes in the observed patterns of secondary structure or the formation of specific contacts are observed for both simulations.

Do Residues Pack Correctly?

A major problem in the calculation of protein structure involves the interaction with solvent. Novotny et al.[18] showed that in vacuo molecular mechanics applied to a completely incorrect starting conformation was able to produce a minimally altered structure with a potential energy comparable to the true structure. While their final structure had no bad van der Waals' contacts or strained covalent interactions, the distribution of polar and nonpolar residues at the protein surface and interior was obviously skewed and thus affected the electrostatic contribution to the enthalpy when solvent was taken into account. A correct globular structure

should be expected to bury a significant fraction of its hydrophobic residues and to expose most charged residues. To determine if the model behaved in this way, an algorithm developed by Ponder and Richards[19] was used to define which residues are buried for each structure produced by the simulation. The Kite–Doolitle index[20] was used to calculate the hydropathy of buried and exposed residues. The average hydropathy of buried residues was 0.69, significantly higher (indicating greater hydrophobicity) than that for the nonburied residues (0.10). The tendency to optimize hydrophobic forces by burying nonpolar residues and excluding highly polar residues appears to be well satisfied for these structures. The corresponding values for the native protein, 0.40 for buried residues and 0.33 for exposed residues, indicated less of a tendency for a hydrophobic core to form in the native protein than in the simulated models. Crambin has unusual water solubility for a protein and forms crystals which are surprisingly nonhydrated.[18] The average hydropathy of buried and exposed residues in several other proteins was calculated to determine if the typical model structure or the native crambin structure is unusual with respect to partitioning of hydrophobic and hydrophillic residues. The average difference between buried and exposed residues in bovine pancreatic trypsin inhibitor (5PTI), cobra venom toxin (1CTX), staphylococcal nuclease (2SNS), and α-lytic protease (2ALP) was calulated using the PDB structures for these proteins (data not shown). The mean difference in hydropathy between buried and exposed residues for these test proteins was 1.30, significantly higher than that for the both native and model crambins. It appears that the simulation is attempting to force the formation of a hydrophobic core for a protein which does not require it. The potentials used for the simulation may not be appropriate for a protein such as crambin whose structural properties differ significantly from those of the basis set which originally defined the potentials. The fact that several structures could be found with a calculated energy lower than that of the native may be due to the incorrect choice of potentials for a protein that does not enormously favor the formation of a hydrophobic core. The stability of crambin may be due largely to disulfide bond formation rather than optimization of the hydrophobic effect. The ability of the model to simulate disulfide bond formation will be discussed later.

Formation of Specific Contacts

The time during the simulation when the final contacts first occur was tabulated for each structure and used to calculate the average age of each contact. The largest contact region, that between the β-sheet strands, was usually formed toward the end of the simulation. This is not surprising since the two helices to which the strands are attached must move together before allowing the strands to be closely paired in an antiparallel sheet. The oldest contacts, as expected, form between residues close to and on opposite sides of each of the turn regions. These contacts are those mostly likely to be associated with a change in conformation which will bring large numbers of residues in contact with one another. The average contact map shows many features of the native, e.g., specific contacts between the two β-sheets, helix–helix pairing at well-defined sites, disulfide formation between the N- and C-termini, etc. The map is inaccurate in that the position of the unordered C-terminus is averaged somewhat over several positions. This may be expected since the lack of assigned secondary structure for the 11 C-terminal residues allows them considerably more conformational flexibility than for the sequences of assigned secondary structure.

Effect of Disulfide Formation on the Folding Pathway

The potential for the interaction between two cysteine residues is about twice as deep as the average potential. One way that inaccuracies in the average contact map could occur would be if the formation of a disulfide locked the conformation of intervening residues and prevented their proper minimization. The formation of native and nonnative disulfides was compared for the 100 structures produced by simulated annealing. The native protein is cross-linked by disulfides between residues 16–26, 4–32, and 3–40. Defined in terms of α-carbon separation (less than 6 Å), a native disulfide also exists between residues 3 and 32. During the folding simulation, the native disulfides occurred more often than any of the nonnative disulfides although they occurred only 0.32 times on average. Nonnative disulfides occurred much less frequently, an average 0.035 times. The average age of all disulfide contacts (15.8% of the total simulation length) is slightly more than the average for all nondisulfide contacts (13.5%). If disulfide formation acts as a steep local minimum to prevent the continued searching of other conformations, one would expect the disulfide contacts to be among the last contacts formed, significantly younger than the nondisulfides.

One way to test the role of disulfides in the simulated folding process is to alter the Cys-Cys potential. By substituting the potential for Cys-Cys residues with one calculated for all residues (generalized potential), the number of both native (0.138/structure) and nonnative disulfides (0.015/structure) was significantly decreased. Of these native disulfides 80% were due to links between residues 3–32 and 4–32, likely due to the natural tendency for β-sheets to pair. The Cys-Cys potential was also replaced by one based on the distribution of disulfide distances rather than Cys-Cys distances. Since no

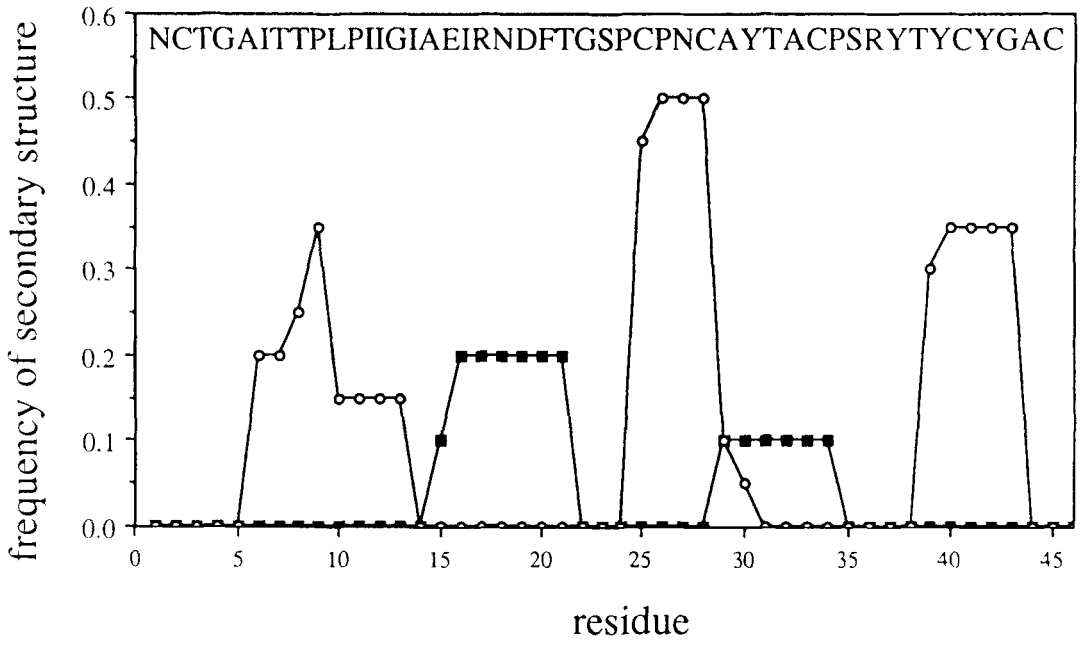

Fig. 8. Results with a shuffled sequence. The above randomly generated sequence, with the same composition as crambin but no homology to it, was used to test the sequence dependence of the modeling. The formation of secondary structure was calculated as described in Figure 6 (open circles indicate α-helical structure while solid squares indicate β-sheet structure).

disulfide bonds are formed at 15–16 Å α-carbon separation, this potential was centered at an arbitrary low level (−20 kT) with zero energy for distances outside of the distribution. Replacing native disulfides (3–40,4–32,16–26) with a disulfide-specific potential effectively selected for the formation of native disulfides over nonnative (0.50 vs. 0.05), although their number did not increase dramatically. Explicit disulfides did adjust the relative occurrence of certain native contacts, increasing the fraction of structures with contacts in the A_1–A_2 and B_1–R regions, but did little change the overall pattern of contact formation. Deepening the disulfide potential for native disulfide pairs apparently does not lead to premature freezing of the conformation as judged by the lack of nonnative contacts.

Structures Produced by a Shuffled Sequence

The structure of a protein is obviously sensitive to its sequence. A model which generates the same structure regardless of sequence is thus certainly incorrect. A simple test of the current model is to randomly shuffle the crambin sequence, maintaining composition, and to repeat the simulation. The shuffled sequence (Fig. 8) was initially annealed with no secondary structure assigned. Secondary structure formed in a well-defined way, but the pattern of helices and sheets was completely different from that of the native sequence. With this secondary structure assigned, the final average contact map is entirely different from that produced by the native sequence. Even with the correct native secondary structure assignment, the contact map produced by the shuffled sequence did not resemble the native sequence-based model. These observations support the hypothesis that the model using the native sequence is mimicking crambin's structure in a sequence-dependent way.

Results for Other Proteins

Preliminary work has been done to simulate the folding of proteins other than crambin. Bovine pancreatic trypsin inhibitor has been used as a standard for many protein folding schemes because of its small size and experimental information on its folding pathways.[21] One of the structures produced by simulated annealing with the current model has a DME deviation of 4.5 Å from the crystal structure. Levitt and Warshel[1] and Kuntz et al.[3] have used other approaches to simulate the folding of BPTI. The best structures generated by their methods have a DME deviation of 5.3 Å and 4.7 Å, respectively, slightly worse than that generated by the current model using simulated annealing. While this improvement is small, it is important to note the fundamental differences between the starting assumptions of other models and our own. As mentioned before, the Levitt–Warshel model assumes certain properties of the starting conformation and the energetics that would force the formation of a BPTI-like structure. Kuntz et al. use a model with several adjustable parameters which are fitted to optimize the agreement to the BPTI crystal struc-

ture. In addition, they assume that the three disulfides found in the crystal structure are known constraints before the minimization. In contrast, our model is parameterized completely by the database of known crystal structures and assumes nothing about the BPTI structure beyond the sequence of amino acids. It is therefore quite surprising that the current model appears slightly better than other models, which include BPTI-specific constraints. The improvement suggests that the specific constraints imposed in these earlier models can be replaced in a generalizable way by potentials derived from a large database of observations. Subsequent work shall determine whether the annealing procedure yields intermediate structures which are similar to those inferred experimentally from data on disulfide formation. If such intermediate structures exist, one may conclude that the annealing process is reproducing the dynamics of the folding process, independent of its ability to converge to the native minimum structure.

CONCLUSIONS

A simplified representation of proteins is presented in which the conformation of the backbone and side chains is specified by the dihedral angels for each residue. An empirical energy is calculated using distance-dependent potentials between each pair of amino acids. These potentials have been derived from the distribution of pairwise distances observed in known crystal structures. Simulated annealing is used as a refinement technique to sample many different conformations and minimize the energy of the model structure. Several results have been obtained using the crambin sequence to test the model:

1. Starting from random conformations, secondary structure forms in the model peptide where it occurs in the true structure.
2. The formation of secondary structure is sequence specific and depends upon the context of the sequence (i.e., is influenced by long-range interactions).
3. With the secondary structure assigned, α-helices and β-strands associate as they do in the true structure.
4. Certain pairs of cysteines, corresponding to native disulfides, associate much more frequently than other cysteine pairs. This pairing is partially driven by the folding of the rest of the protein, and partially driven by the strong attractive potential between two cysteines.
5. The empirical potentials favor the formation of a hydrophobic core of residues. This partitioning is imposed during the crambin simulation, even though the true crambin structure lacks a hydrophobic core.

One problem with the empirical approach used by this model is that the true energetics which drive protein folding are deeply embedded in the amino acid pair potentials. One might ask, for instance, why does a helix form in crambin between residues 22 and 29? The reason cannot be attributed by this model to particular hydrogen bonds, backbone conformational preferences, etc., since the simulation depends on the entire collection of empirical potentials assigned to these and other amino acids. Since, however, minimization using the empirical potentials successfully predicts a helix in this region, the problem has been reduced to explaining how specific interactions within the data base of structures give rise to the potentials in the first place.

ACKNOWLEDGMENTS

We thank Dr. Robert Fox and Dr. David Agard for helpful discussions and computational resources and the San Diego Supercomputer Center for Cray X-MP CPU time donated for the initial studies. CW gratefully acknowledges support from the Hertz Foundation.

REFERENCES

1. Levitt, M., Warshel, A. Computer simulation of protein folding. Nature (London) 253:694–698, 1975.
2. Tanaka, S., Scheraga, H.A. Medium- and long-range interactions parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules. 9(6):945–950, 1976.
3. Kuntz, I.D., Crippen, G.M., Kollman, P.A., Kimelman, D. Calculation of protein tertiary structure. J. Mol. Biol. 106: 983–994, 1976.
4. Hagler, A.T., Honig, B. On the formation of protein tertiary structure on a computer. Proc. Natl. Acad. Sci. U.S.A. 75(2):554–558, 1978.
5. Go, N., Taketomi, H. Respective roles of short- and long-range interactions in protein folding. Proc. Natl. Acad. Sci. U.S.A. 75(2):559–563, 1978.
6. Ramachandran, G.N., Sasisekharan, V. Conformation of polypeptides and proteins. Adv. Protein Chem. 23:283–437, 1968.
7. Crippen, G.M., Viswanadhan, V.N. Sidechain and backbone potential function for conformational analysis of proteins. Int. J. Peptide Protein Res. 25:487–509, 1985.
8. Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P. Optimization by simulated annealing. Science 220:671–680, 1983.
9. Nigles, M., Gronenborn, A.M., Brunger, A.T., Clore, G.M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance constraints. Application to crambin, potato carboxypeptidase inhibitor, and barley serine proteinase inhibitor 2. Protein Engineer. 2:27–38, 1988.
10. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092, 1953.
11. Hendrickson, W.A., Teeter, M.M. Structure of the hydrophobic protein crambin determined directly from the anomolous scattering of sulphur. Nature (London) 290:107–113, 1981.
12. Teeter, M.M. Water structure of a hydrophobic protein at atomic resolution. Pentagon rings of water molecules in crystals of crambin. Proc. Natl. Acad. Sci. U.S.A. 81:6014–6018, 1984.
13. Cohen, F.E., Sternberg, M.J.E. On the prediction of protein structure: The significance of the root mean square deviation. J. Mol. Biol. 138:321–333, 1980.
14. Havel, T.F., Kuntz, I.D., Crippen, G.M. The theory and practice of distance geometry. Bull. Math. Biol. 45:665–720, 1983.

15. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochemistry 13(2):222–245, 1974.
16. Cohen, F.E., Abarbanal, R.M., Kuntz, I.D., Fletterick, R.J. Secondary structure assignment for alpha/beta proteins by a combinatorial approach. Biochemistry 22:4894–4904, 1983.
17. Schulz, G.E., Schirmer, R.H. "Principles of Protein Structure." New York: Springer-Verlag, 1979.
18. Novotny, J., Bruccoleri, R., Karplus, M. An analysis of incorrectly folded protein models: implications for structure predictions. J. Mol. Biol. 177:787–818, 1984.
19. Ponder, J.W., Richards, F.M. Tertiary template for proteins-use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193: 775–791, 1987.
20. Kite, J., Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
21. Creighton, T.E. Experimental studies of protein folding and unfolding. Prog. Biophys. Mol. Biol. 33:231–297, 1978.