PROTEINS: Structure, Function, and Genetics 14:409–420 (1992)

# Folding Protein α-Carbon Chains Into Compact Forms by Monte Carlo Methods

**David G. Covell**
*Frederick Cancer Research and Development Center, National Cancer Institute, Program Resources, DynCorp Inc., Frederick, Maryland 21702*

***ABSTRACT*** **A method is presented for generating folded chains of specific amino acid sequences on a simple cubic lattice. Monte Carlo simulations are used to transform extended geometries of simplified α-carbon chains for eight small monomeric globular proteins into folded states. Permitted chain transitions are limited to a few types of moves, all restricted to occur on the lattice. Crude residue–residue potentials derived from statistical structure data are used to describe the energies for each conformer. The low resolution structures obtained by this procedure contain many of the correct gross features of the native folded architectures with respect to average residue energy per nonbonded contact, segment density, and location of surface loops and disulfide pairs. Rms deviations between these and the native X-ray structures and percentage of native long-range contacts found in these final folded structures are 7.6±0.7 Å and 48±3%, respectively. This procedure can be useful for predicting approximate tertiary interactions from amino acid sequence.**



**Key words: lattice models, folded proteins, compact states**

## INTRODUCTION

A strong body of accumulated evidence indicates that protein structure and stability are the resultant of many complex forces that favor and oppose the folded native state. Although the precise balance between these forces remains unknown, two are thought to dominate in their contributions to folding. The seminal work of Kauzmann proposed that the transition of small single-domain globular proteins from the denatured to the native state is driven predominantly by the so called *hydrophobic collapse,*[1,2] a term referring to the transfer of nonpolar solutes into aqueous solution. The dominant force opposing this collapse is thought to arise from loss of *configurational entropy* due to volume exclusion constraints of the folded state.[3] This latter term arises from the difficulty of packing a flexible chain into a confined space. Theoretical models capable of incorporating the precise details of these dominant forces into a form sufficient for predicting native states is not yet computationally feasible;[4–6] less detailed models have not been sufficiently explored and might offer a practical alternative.

This work resembles in some ways approaches proposed by others in the use of an extremely simplified model of small monomeric globular proteins[7–16] to produce folded conformations. The novelty of the approach described here lies in the method of calculation. Lattice-restricted simulations with excluded volume considerations are used to move an extended α-carbon backbone chain through transitions to a low energy state. Rules defining the energies of each conformation are based on hydrophobic interactions derived from residue–residue contact statistics.[17] The structures obtained are examined for their agreement with nonbonded residue contacts found in native structures and for their similarity to the native folded topology (see Kuntz et al.[9] for a discussion of criteria for comparisons).

A simplified model of native globular proteins can be imagined as a string of beads tightly packed into a dense globule[18] with variable bead sizes specified by their amino acid types. Residue interactions depend directly on other spatially close residues. Within the core of the globule there are preferred nearest-neighbor packings; beads with strongly hydrophobic character prefer similar adjacent beads rather than beads with less hydrophobic or more hydrophilic character.[1,17,19] The opposite is true on the outside of the protein; the preference is for hydrophilic and weakly hydrophobic residues. These general features are thought to result from rules that favor burial of nonpolar surfaces in the process of collapse to a denser folded state.[20–22] An important feature of this collapse, that distinguishes it from collapsed homopolymers, is the segment ordering which substantially, but imperfectly, segregates nonpolar residues into the core.[3,17,23,24]

The collapse of a denatured protein to its native

Received October 15, 1991; revision accepted March 5, 1992.
Address reprint requests to Dr. David G. Covell, Frederick Cancer Research and Development Center, National Cancer Institute, Program Resources, DynCorp Inc., Building 430, Frederick, MD 21702.

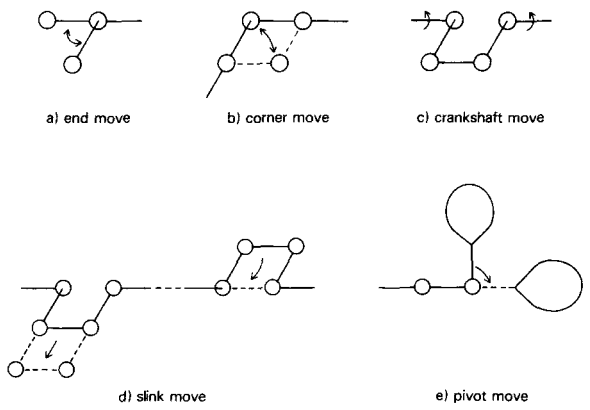


Fig. 1. Allowed chain transitions **(a–e)** for the lattice model. Transitions are permitted only to previously unoccupied lattice sites.

form occurs in a highly cooperative phase transition.[25,26] Models of this transition must take into account the exceptional cooperativity between chain elements of the native state. Presumably the cooperativity must increase as chain density increases. One difficulty in describing this transition lies in the correct formulation of constraints imposed on the moving chain by its interactions with solvent, ions, and other parts of the chain, both in the mean field sense as well as for specific interactions. Tools to explore these effects can be found in molecular dynamics simulations[27] where classical Newtonian equations of motion are solved to determine atom movements in response to interaction forces. Other approaches have focused on theories of molecular relaxation to explore polymer chains in condensed states.[28] These latter studies attempt to interpret stress–relaxation and dielectric–relaxation data of polymer systems. The procedure described here bears some similarities to both of these approaches.

## LATTICE MODEL

There are two essential considerations in the model: chain geometry and chain energy. The geometry consists of a single linear chain of elements on a regular, simple cubic lattice. The unit spacing of the lattice is fixed at 3.8 Å, the fixed value of virtual α-carbon bonds observed in real proteins. Chain movement consists of relocating chain elements using a few simple steps. The steps permitted here (see Fig. 1) include (1) *end moves*,[29–31] which relocate chain ends to any available nearest lattice point, (2) *corner* moves,[29–31] where any corner position can be moved to the opposite corner, (3) *crankshaft* moves,[32] where a kink (*cis* arrangement of four consecutive chain elements) is rotated about the axis determined by the position of the outermost chain elements, (4) *slink* moves, where a pair of quartets of *cis* bonded chain elements is modified by extending the central elements of one pair by two bonds and deleting two bonds from the other, and (5) *pivot* moves[33–35] where a contiguous group of chain elements is moved as a rigid body around a central chain element. All moves are taken to be self-avoiding and depend strictly on the availability of unoccupied space into which the transition occurs. Additional volume exterior to that of the chain is thus required for transitions. *End* or *corner* moves involve repositioning one chain element. A *crankshaft* move involves relocation of the two bonded central chain elements of the quartet. A *slink* move involves inserting and deleting, for compensation, two chain elements at different locations in the chain. The number of elements repositioned in a *pivot* move depends on the location of the chain element around which the pivot is made. These four types of transitions provide the chain with local flexibility, useful for improving the details of chain packing, and large scale motions, which permit a broad exploration of conformational space.[36] Chain rearrangements always preserve the identities of individual chain elements as well as their bond lengths of 3.8 Å.

Chain energies are described by residue potentials. These potentials are divided into three parts: an attractive, a surface, and a packing term:

$$E_{\text{total}} = E_{\text{attractive}} + E_{\text{surface}} + E_{\text{packing}}.$$

The attractive interactions between spatially close residues are determined as the sum of the energies of all nonbonded contact pairs, $e_{ij}$, within the distance $d = 7.5$ Å. This cutoff distance is consistent with nearest neighbor packing of α-carbons found in small globular proteins[7]

$$E_{\text{attract}} = \sum_{\text{contacts}} e_{ij}$$

where $e_{ij}$ represents the statistically derived energy difference between $i$-$j$ residue pairs and $i$-solvent plus $j$-solvent pairs. The strengths of these contact energies occur in decreasing order of their contribution to protein stability so that $e_{\text{hydrophobic,hydrophobic}}$ individually contributes the most, $e_{\text{hydrophilic,hydrophobic}}$ intermediate, and $e_{\text{hydrophilic,hydrophilic}}$ the least. Additional details of the derivation of $e_{ij}$ can be found in Table 5 of Miyazawa and Jernigan.[17] Covell and Jernigan[7] demonstrated the utility of these contact energies for distinguishing native from nonnative backbone conformations in confined spaces, providing confidence that relative ordering of the importance of intraresidue contacts is possible. This approach is also strongly motivated by conclusions that the all-or-none character of folding transitions is likely to arise from the specificity of long-range contacts.[37,38]

A crude energy term related to surface exposure is defined as follows:

$$E_{\text{surface}} = -G/s^2$$

where $s^2$ is the radius of gyration squared and $G$ is a scalar constant whose value was set so $E_{surface}$ contributes ~10% of $E_{total}$. The addition of this term to $E_{total}$ favors the occurrence of more compact conformations. $E_{surface}$ is most effective at the beginning of each simulation where more extended chains exist and least effective at the end of each simulation where chains are most compact.

A severe criticism of using a simple cubic lattice to model an α-carbon backbone is that it permits unrealistically high chain packing densities. McGregor and Cohen[39] estimate simple cubic chain packing densities on a lattice with 3.8 Å unit spacing to be ~30% greater than native densities. An additional but related criticism is that the number of near neighbors for a simple cubic lattice does not closely relate to that found in globular proteins. Miyazawa and Jernigan[17] estimate an average of six nearest neighbors for all amino acid types within 6.5 Å of a sphere positioned at the centroid of each side chain. A similar number of neighbors is found for α-carbon spacings at 7.5 Å.[7] At this distance, however, the simple cubic lattice has 22 possible neighbors. In an effort to make the simple cubic lattice used here more compatible with the packing conditions observed by Miyazawa and Jernigan, and therefore more compatible with the residue–residue energy potentials derived in their study, the following strategy is used. For each residue's contribution to $E_{attractive}$, only the six best residue–residue contact energies from the total set of nearest neighbors ($=k$) are used in the summation:

$$e_{residue} = \sum_{j=\min(6,k)} e_{residue,j}.$$

In addition, to compensate for overly compact chains on the cubic lattice, a large penalty, $E_{packing}$, is added to the total energy when chain density is too high. Chain densities are estimated by counting the total number of nonbonded interactions for each chain. The total number of residue–residue interactions in a chain of length $N$ confined to a volume $V$ is assumed to be proportional to $N^2/V$.[40] Calculations for the set of proteins considered indicate that for a nonbonded cutoff distance of 7.5 Å and estimates of volume from radius of gyration, a value of 200 Å$^3$ is a reasonable proportionality constant. A large positive constant was thus added to $E_{total}$ for conformations whose total number of contacts exceeds $200N^2/V$. Clearly this constant depends on local chain densities and composition, but this is a first-order approximation used to simply prevent too high an overall chain density.

Undoubtedly other more detailed terms in atomic interaction energies contribute to the selection of native conformations.[41] The main thrust of this model deals with a broader exploration of conformational space with these residue–residue contacts than is possible for more detailed atomic interaction potentials.

## COMPUTER SIMULATIONS

Calculations are performed on a three-dimensional simple cubic lattice in a box comprising 12 sites in each direction with a unit spacing of 3.8 Å. Chain elements cannot extend beyond the boundaries of this cube. Lattice models simply serve to discretize conformational space and to facilitate rapid calculations of large numbers of conformations.[7] Chain collapse is achieved by simulation using an energy-based sampling algorithm. All chain rearrangements are restricted to the lattice. Each simulation starts from a randomly generated extended chain with $s^2$ in the range of 3 to 5 times that of the native protein. This ratio falls within the range of values estimated for random coil conformations.* Initial RMS deviations from native forms are typically in the range of 15–25 Å. At each simulation step, all possible *end, corner, crankshaft,* and *slink* transitions, and 30 randomly chosen *pivot* transitions are determined for the current chain conformation. The *pivot* moves are chosen by randomly selecting a chain element and randomly selecting, from the set of moves possible for a self avoiding chain, a move around this pivot point. $E_{total}$ is determined for each conformation in the sample. A uniformly distributed random number is used to select the next conformation from among this set of possible conformations based on its Boltzmann weight relative to all in the sample. This scheme resembles the procedure of Meirovitch[44] where the next move is selected from the complete set of transition moves. Choosing from the complete set of moves may be a more efficient way to sample conformational space than the conventional Metropolis method.[12,42] The next possible chains having the lowest $E_{total}$ are chosen with the greatest probability at each simulation step. The net effect is to reduce, on average, $E_{total}$ to a low value. No comparison is made between the energy of the newly selected conformation and that of the previous conformation, further distinguishing this approach from that of Metropolis. This method of conformational searching is not so likely to be trapped in a local energy minimum because there is always a finite chance of passing to less favorable energy forms.

Previous efforts to "melt" extended chains to compact forms have been criticized on the conjecture that starting conformations are simply high energy distorted forms of the native state. Methods to address this criticism are not clear. The approach used here is to construct the coordinates of the final

*From the relation of Sanchez,[43] $s^2/\langle s^2\rangle_o = (\sigma_o/\sigma)^{2/3}$, where the subscript o refers to the random coil state and σ is segment density. With a packing density of $\sigma = 0.74$ for ideal spheres, $\sigma_o = (19/27N)^{1/2}$, and a protein of length $N = 70$, $s^2/\langle s^2\rangle_o \sim 4$.

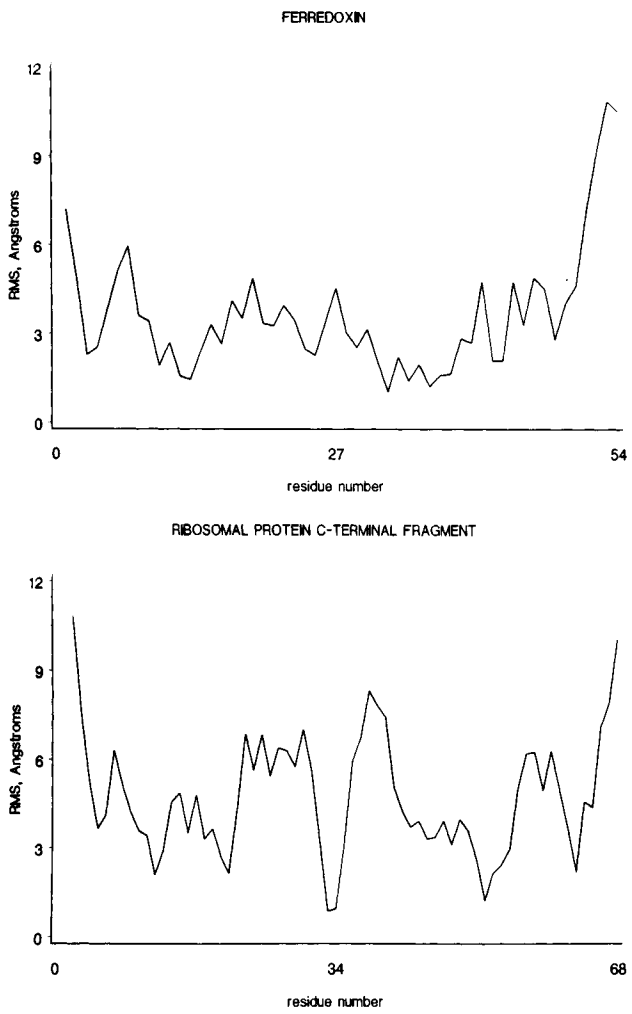


Fig. 2. Rms versus residue number for **(Top)** ferredoxin (1FDX) and **(Bottom)** ribosomal protein C-terminal fragment (1CTF). Values represent each residue's contribution to the total rms deviation between the native and folded forms.

**TABLE I. Measures of Goodness of Fit***

| Protein† | Length | rms (Å‡) | Native contacts (%) |
|---|---|---|---|
| 1CRN | 46 | 7.5 (6.8–8.7) | 53 |
| 5RXN | 52 | 7.3 (6.9–8.6) | 47 |
| 1FDX | 54 | 6.1 (5.8–8.2) | 49 |
| 1OVO | 56 | 7.5 (7.0–8.8) | 44 |
| 5PTI | 58 | 7.6 (7.4–8.1) | 46 |
| 1CTF | 68 | 8.5 (8.1–10.0) | 46 |
| 1HOE | 74 | 7.9 (5.8–10.1) | 49 |
| 1UBQ | 76 | 8.0 (7.8–9.2) | 46 |
| <ave> | | 7.6 ± 0.7 | 48 ± 3 |

*Model simulations are done on a Cray XMP24. Each simulation step requires ~1 CPU second. Final folded structures for each starting conformation are determined as the average of the minimum energy forms obtained for five simulations. Three thousand simulation steps were done for each starting conformation.

†Native protein coordinates were obtained from the Brookhaven Protein Data Bank.[61] 1CRN, crambin; 5RXN, rubredoxin; 1FDX, ferredoxin; 1OVO, ovomucoid third domain; 5PTI, pancreatic trypsin inhibitor; 1CTF, ribosomal protein C-terminal fragment; 1HOE, α-amylase inhibitor; 1UBQ, ubiquitin.

‡Range of rms deviations for the 5 simulations are shown in parentheses.

folded form as the arithmetic average of the coordinate positions of the lowest energy chains found in five simulations. The selection of five runs is made because fewer runs may not sufficiently sample a range of possible low energy conformations. Ideally many more simulations would be better; five is selected as a compromise.

Each simulation begins with a randomly constructed chain. All proteins studied use the same set of five starting conformations; proteins smaller in length than the longest protein studied ($N = 76$) are simply truncated from the carboxyl-termini. Using the same set of starting conformations ensures that each protein is treated identically as far as the simulation conditions are concerned and that the different final folded result is due only to differences between each protein's amino acid sequence. The coordinate positions of the final folded forms are the arithmetic average of coordinates from the minimum energy conformation obtained in each simulation run [$<x> = (x_1 + x_2 + x_3 + x_4 + x_5)/5$]. An iterative procedure is used to obtain the coordinates used for averaging. First the rms deviation between the five chains is calculated. Second, the coordinates of the most closely matched pair are averaged to define a reference state. Subsequently, among the remaining chains, the one with the lowest rms deviation from the current reference chain is matched to the reference chain and then included in the average. This process is repeated until all chains have been considered. Using these averaged coordinates as positional constraints, a polypeptide backbone is constructed with the procedure of Correa.[45] The latter step is necessary to regularize the averaged α-carbon coordinates in terms of proper bond length and to partially compensate for the coordinate compression due to averaging.

Evaluation of the folded topologies obtained using the procedures outlined above requires a means to distinguish between features of the folded conformers due to amino acid sequence from those due simply to chain compaction. Reference states for making such comparisons are generated by repeating the above procedures for a homophobic (poly-Leu) sequence. Simulations for chains of 52 and 58 amino acid residues were completed for comparison with the results for 5RXN and 5PTI, respectively. These two proteins were selected because (1) the fold for 5RXN appears to be somewhat simpler, as evidenced by previous successes at folding to its correct

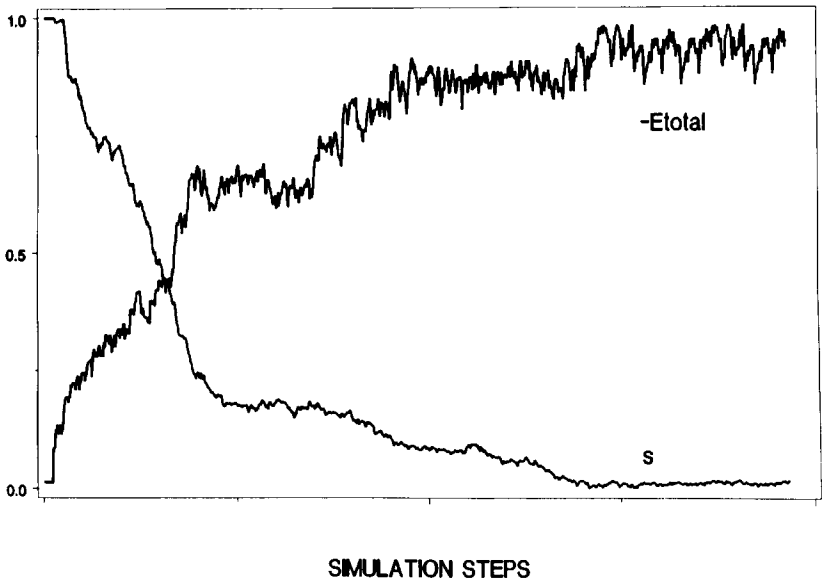


Fig. 3. Radius of gyration, $\sqrt{s^2}$, and $-E_{total}$ versus simulation steps for 1CTF. Ordinate values are scaled from zero to one, for lowest and highest observed values.

topology,[9,11] and may be due simply to chain compaction, and (2) the folding of 5PTI has been more difficult to obtain and may better represent a case where amino acid sequence is more critical.

## RESULTS

Eight small monomeric globular proteins are examined using the proposed model. Their lengths ranged from 46 amino acids for crambin (1CRN) to 76 amino acid residues for ubiquitin (1UBQ). Measures of goodness of fit between native and final folded forms will include rms deviations, distance maps[46,47] and the correctness of turn locations and disulfide pairings.

Rms deviations for the folded forms from their native structures are largest at 8.5 Å for 1CTF and smallest at 6.1 Å for 1FDX. See Table I for details. These results compare favorably with rms errors obtained by Kuntz et al.,[9] Levitt and Warshel,[11,48] and Skolnick and Kolinski.[49] These published cases were obtained by using either molecule-specific pushing and pulling potentials, forcing correct iron/sulfur bonds, or favoring specific turn locations, respectively. The results for the eight proteins studied indicate only fair agreement with the crystal structure coordinates (<rms> = 7.6±0.7 Å). "Ideal" simple cubic lattice models of α-carbon backbones fit their native counterparts with an rms deviation in the range of 3.5–4.5 Å. Calculations by Correa[45] indicate that rms values for α-carbons in the range of 2–3 Å from native may be sufficient to construct useful all-atom structures. Inspection of the quality of fit between folded and native structures indicates that residues showing the largest deviation are often found near chain termini (see Fig. 2). Goodness of rms fit did not appear to depend on protein length or type of secondary structure found in these proteins.

A representative plot of radius of gyration, $\sqrt{s^2}$, and $E_{total}$ during a simulation process is shown in Figure 3. These results consistently follow the same pattern; $\sqrt{s^2}$ and $E_{total}$ rapidly decrease to plateaus early in the process. Once the initial plateau is reached, additional decreases to lower plateaus occur until the final level is reached. These plateaus may be related to the folding processes of random condensation followed by chain rearrangement proposed by Dill.[50] The first appearance of a plateau consists of conformations with a radius of gyration around 50% greater than the final value, while the average radius of gyration for the plateau immediately before the final value is in the range of 5 to 10% greater. Kuntz et al.[9] made similar observations for rubredoxin using a different approach. These decreases in radius of gyration suggest that chain collapse occurs as a succession of internal chain rearrangements which increase compactness and at the same time improve nonbonded contacts. Calculations by Shakhnovitch et al.[51] and results of Kuwajima[52] suggest that the freedom afforded in a somewhat expanded state may be critical to facilitate rearrangements necessary before collapse to the native structure can occur.

The gross physical features of the folded forms compare favorably with their native structures. The radius of gyration and α-carbon density for all proteins studied are within 7% of their native values.

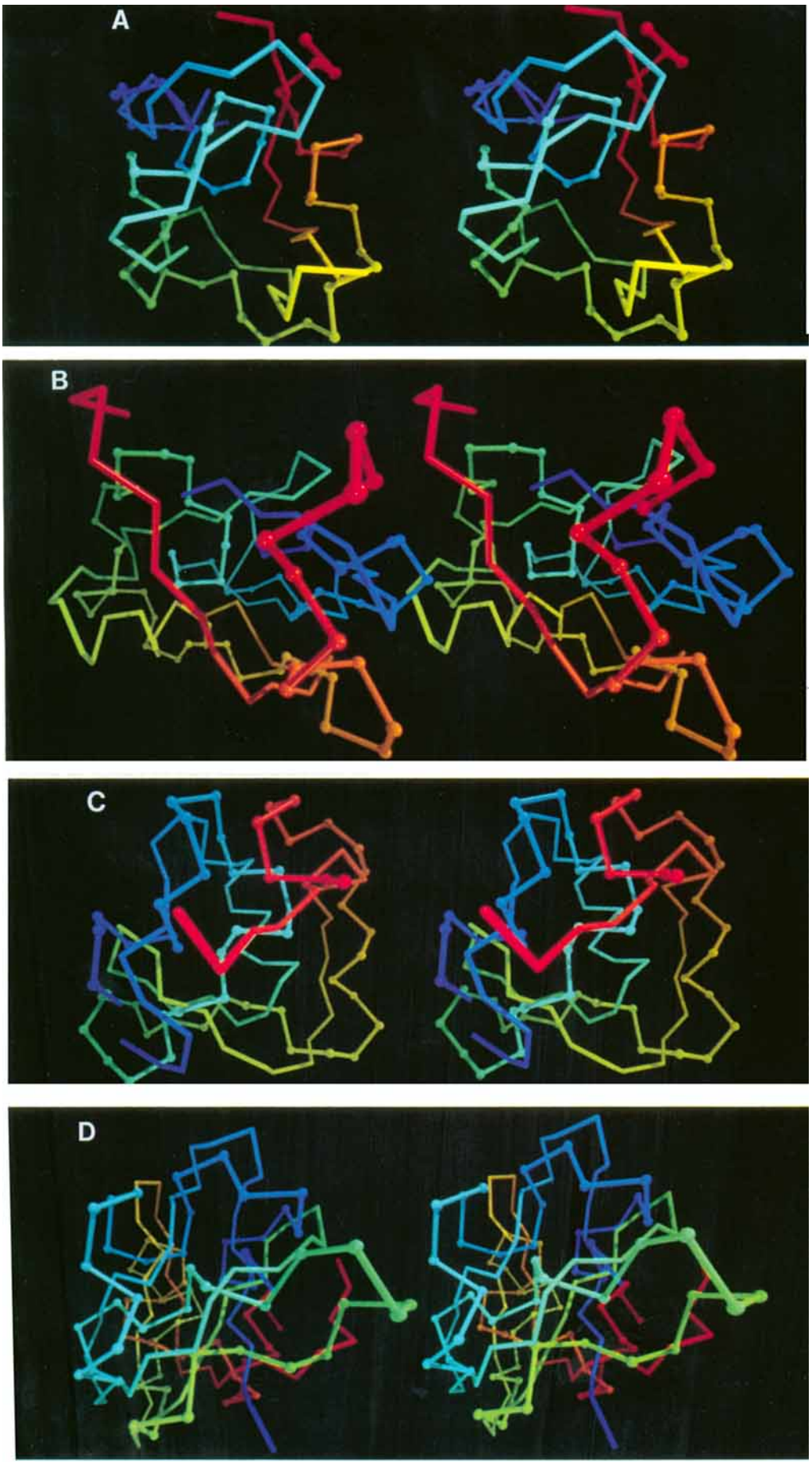


Fig. 4.

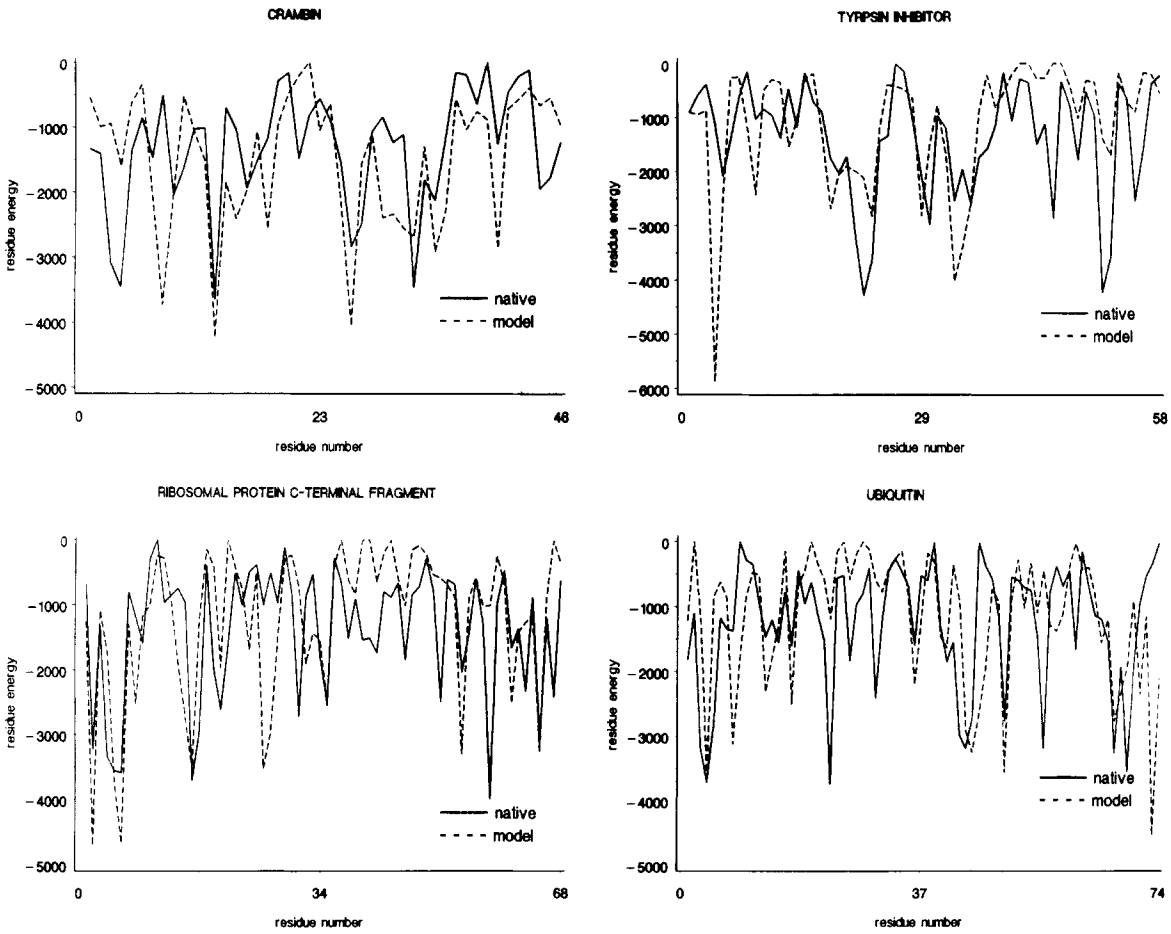


Fig. 5. Residue energies for the native (solid line) and final folded form (dashed line) as a function of residue number for crambin (1CRN), trypsin inhibitor (5PTI), ribosomal protein C-terminal fragment (1CTF), and ubiquitin (1UBQ). Energy for residue $i$ is calculated as $\Sigma e_{ij}$ for all residues $j$ within 7.5 Å.

These results suggest similarities of chain compaction for the two forms. Chain stiffness, resulting from the small set of allowed chain transitions, and limiting the number of residue–residue contacts in $E_{\text{total}}$ must contribute to this result. This result is in contract with those of Skolnick and Kolinski;[49] they found folded forms significantly more compact than the native state using a different procedure.

Backbone tracings of the folded and native structures for rubredoxin (5RXN), ovomucoid 3rd domain (1OVO), α-amylase inhibitor (1HOE), and ubiquitin (1UBQ) are shown in Figure 4. These images demonstrate the overall conformational similarities between the two forms. In all cases the core region of the final folded form is topologically correct; topological agreement here indicates that either conformer could be moved toward the other using only local adjustments that do not permit chains crossing through one another or substantial local unfolding. Within the core region there is close packing of important hydrophobic residues. This fact is further illustrated in Table I, where an average of 48% of the nonbonded contacts found in the native structure was also found in the folded structure. An example of segment packing is seen in Figure 5 where contact energies are plotted for each residue in the native and folded state of crambin (1CRN), trypsin inhibitor (5PTI), ribosomal protein C-terminal fragment (1CTF), and ubiquitin (1UBQ). Generally good agreement is observed between the magnitude of interaction energies in the native and folded forms for each residue in the sequence ($r = 0.8$, $P < 0.05$). A strong correlation was also found between $E_{\text{total}}$ for native and final folded structures ($r = 0.94$, $P < 0.05$). These results indicate that a native-like

Fig. 4. Stereo images of the final folded form, shown with expanded spheres at each α-carbon position, and native α-carbon backbone for (**A**) rubredoxin (5RXN), (**B**) ovomucoid 3rd domain (1OVO), (**C**) α-amylase inhibitor (1HOE), and (**D**) ubiquitin (1UBQ). Strands are colored spectrally from amino (red) to carboxyl (blue) terminal. A few residues are removed from the termini of 5RXN, 1OVO, and 1UBQ to improve visualization of core packing.

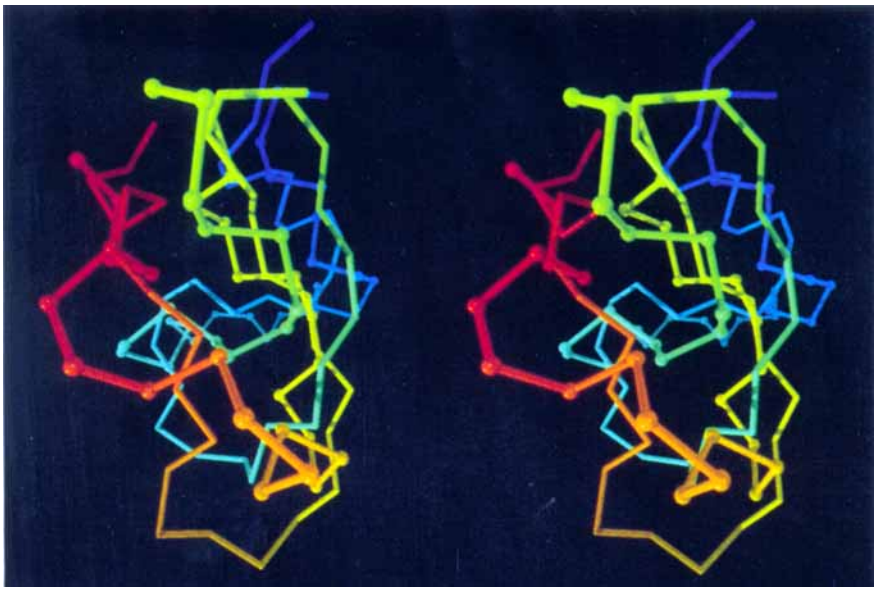

Fig. 6. Stereo images of the final folded form, shown with expanded spheres at each α-carbon position, and native α-carbon backbone for pancreatic trypsin inhibitor (5PTI).

neighborhood of important spatially close residues occurs in the final folded states. As stated earlier, the position of chain termini was often found to deviate from the native structure. Since chain ends typically have the greatest conformational freedom (excluded volume constraints affect chain ends the least) their spatial position was most difficult to define, as it is in crystal structures.

An average interaction energy of 2–3 RT/contact was found for the folded forms. The free energy of each contact can be expected to be greater due to entropic effects. Calculating these effects would require estimation of the entropy of all conformational states. Although such estimates have been made using the Flory–Fisk approximation[50,53] or from computer simulations,[54] these estimates do not directly take into account any specificity for hydrophobic interactions.†

The folded forms generated from the five starting conformations for each protein differ from each other by an average rms deviation of 7.5±1.7 Å. Values of $E_{total}$ for the five low energy conformations for each protein are within 10% of each other. The average coordinate positions used to define the structures reported in Table I have a standard error of the mean for each residue position of 4.3 Å. These results indicate the limits of resolution of this procedure, i.e., they are similar to the intrinsic resolution of the simple cubic lattice used here. Use of this low resolution lattice also contributes to the existence of a family of isoenergetic folded forms rather than one unique folded state. Averaging of coordinate positions for these low energy states yields a topology that is more native-like than each low energy structure. This result is due, in part, to the observation that although any given low energy conformer may have a poor rms deviation from the native structure, rms values for subsegments can be much better. Inclusion of a larger number of simulations can offer the advantage of selecting the most similar sets of conformers for averaging. This should improve the quality of the final folded form. Improvements in resolution can also be expected with a more detailed lattice, such as a face-centered or body-centered cubic lattice,[7] but the computational requirements will be significantly greater, the choice of lattice moves more arbitrary, and generating a folded structure more difficult. There are errors in the energy evaluations, and improvements might come from other treatments of residue–residue potential energies.

Homophobic sequences also fold to compact states. The folded models resulting from chains of length 52 and 58 residues have larger rms deviations when compared to their native states than the results for native sequences; rms values of the folded forms were 2.6 and 3.8 Å greater than the values published in Table I for 5RXN and 5PTI, respectively. A portion of this poorer quality of fit to the native structure is the result of less rms agreement among the five compact states. In both cases the standard error of the mean for averaged coordinate positions is nearly twice that for the reported model folds. It appears that the simulations find states that maxi-

†Although it is, in principle, possible to estimate configurational entropy of each state, following Go et al.,[54] by considering the number of isoenergetic states, such an estimate was not attempted here.

**TABLE II. Comparison of Disulfide Pairings in Native and Model Structures**

| Protein* | SS bond | $D_{native}$ (Å) | $D_{model}$ (Å) | Abs. diff. |
|---|---|---|---|---|
| 2CRN | 3–40 | 4.8 | 13.9 | 9.1 |
| | 4–32 | 4.1 | 10.0 | 5.9 |
| | 16–26 | 5.8 | 3.9 | 1.9 |
| 1OVO | 8–38 | 5.4 | 13.0 | 7.6 |
| | 16–35 | 5.7 | 10.9 | 5.2 |
| | 24–56 | 4.8 | 10.8 | 6.0 |
| 5PTI | 5–55 | 5.7 | 13.9 | 8.2 |
| | 14–38 | 5.7 | 11.9 | 6.2 |
| | 30–51 | 6.4 | 11.0 | 4.6 |
| 1HOE | 11–27 | 5.1 | 9.9 | 3.9 |
| | 45–73 | 6.4 | 4.0 | 2.4 |
| | | | | <5.2 ± 2.1> |

*See Table I footnote for abbreviations.

**TABLE III. Comparison of Turn Locations Between Model and Native Structures**

| Protein* | Turn position† | Rms‡ |
|---|---|---|
| 1CRN | 41–44 | 4.7 ± 3.0 |
| 1CTF | 61–64 | 7.6 ± 2.4 |
| 1HOE | 17–20 | 8.8 ± 2.7 |
| | 37–40 | 9.6 ± 1.8 |
| UBQ | 7–10 | 8.3 ± 3.6 |
| | 18–21 | 9.5 ± 1.2 |
| | 37–40 | 4.6 ± 2.6 |
| | 45–48 | 12.3 ± 0.3 |
| | 51–54 | 5.1 ± 2.5 |
| | 56–59 | 4.6 ± 1.9 |
| | 62–65 | 8.0 ± 1.2 |

*See Table I footnote for abbreviations.
†Turn positions taken from Brookhaven depositor.
‡Rms deviation is average distance between folded and native forms for residues in turns. Structures are rms matched over their entire sequence.

mize the number of nonbonded contacts, regardless of the quality of each interaction. Consequently, there appear to be many isoenergetic conformers for the homophobes with poor rms agreement between them, which when averaged result in a more poorly defined folded state, as reflected in the higher rms values.

The results for homophobes are useful for evaluating the percentage of native long-range contacts listed in Table I. For the two homophobes studied, the percentages of native contacts found in the final folded forms were always 15 to 20% less than found for the native sequences. This means that 20–30% of native-like contacts occur from essentially compactness alone. Further analysis of previous results for all compact states enumerated on a native-like volume[7] supports this result; 20–30% of native-like nonbonded contacts can be found in compact nonnative conformations. At the other extreme, simple cubic lattice models of proteins α-carbon backbones fit native conformations to an rms in the range of 3.5–4.5 Å. These models are quite good at reproducing native chain tracings, with the percentage of native nonbonded contacts being 70–80% of those found in the native state. Therefore, in the window from best (~75%) to worst (~25%) case scenarios, the results of 48% native nonbonded contacts reported in Table I lie in the middle of the expected range. These results indicate that the final folded forms are clearly better than obtained from compaction alone, but do not yet represent the "ideal" case.

Table II summarizes distances between disulfide pairs in native and final folded structures. These results can be used to determine how well the topologies of the final structures bring together correct cysteines. A reasonable agreement between distances separating α-carbons of disulfide pairs was obtained (average rms error = 5.2±2.1 Å). Inspection of these results indicates that most of the larger separation distances involve cysteines near the ends of the molecule where chain position is least well determined. Visual inspection of the cysteine pairs indicates that no intervening strands separated correct disulfide pairings. This suggests that deforming the present conformations by using a stronger cysteine-specific attractive potential could be used to pull appropriate strands closer to the final folded form.

Table III lists the rms deviations between turn locations in the folded and native forms. Turn locations are taken from the Brookhaven depositor. In nearly all cases the positions of turns are similar to that found in the native structures. Not only did they agree reasonably well as far as rms deviation but their relative position on the outside of the molecule was also correct. The one notable exception is for turn 45–48 in 1UBQ, the worst case as far as rms deviation from the native structure, where the folded form has this turn positioned away from the core of the molecule rather than toward the core as in the native state.

Considerable attention has been given to whether computer folded models of 5PTI contain the native features of (1) the 180° twist in the β-sheet involving residues 15–30, and (2) proper threading of strand 30–40 back inside the amino terminal loop.[55,56] Both of these features are evident in the folded form for 5PTI (see Fig. 6). The threading feature is more strongly consistent with the native fold than the twist, where only a partial twist is found. Producing the detailed topology of features like β-sheet twists will require more complete models that preserve chain chirality. Including side chains in the model would presumably improve this approach.

Distance maps for four of the eight proteins studied are shown in Figure 7. These figures reflect the results listed in Table I that about half of the native

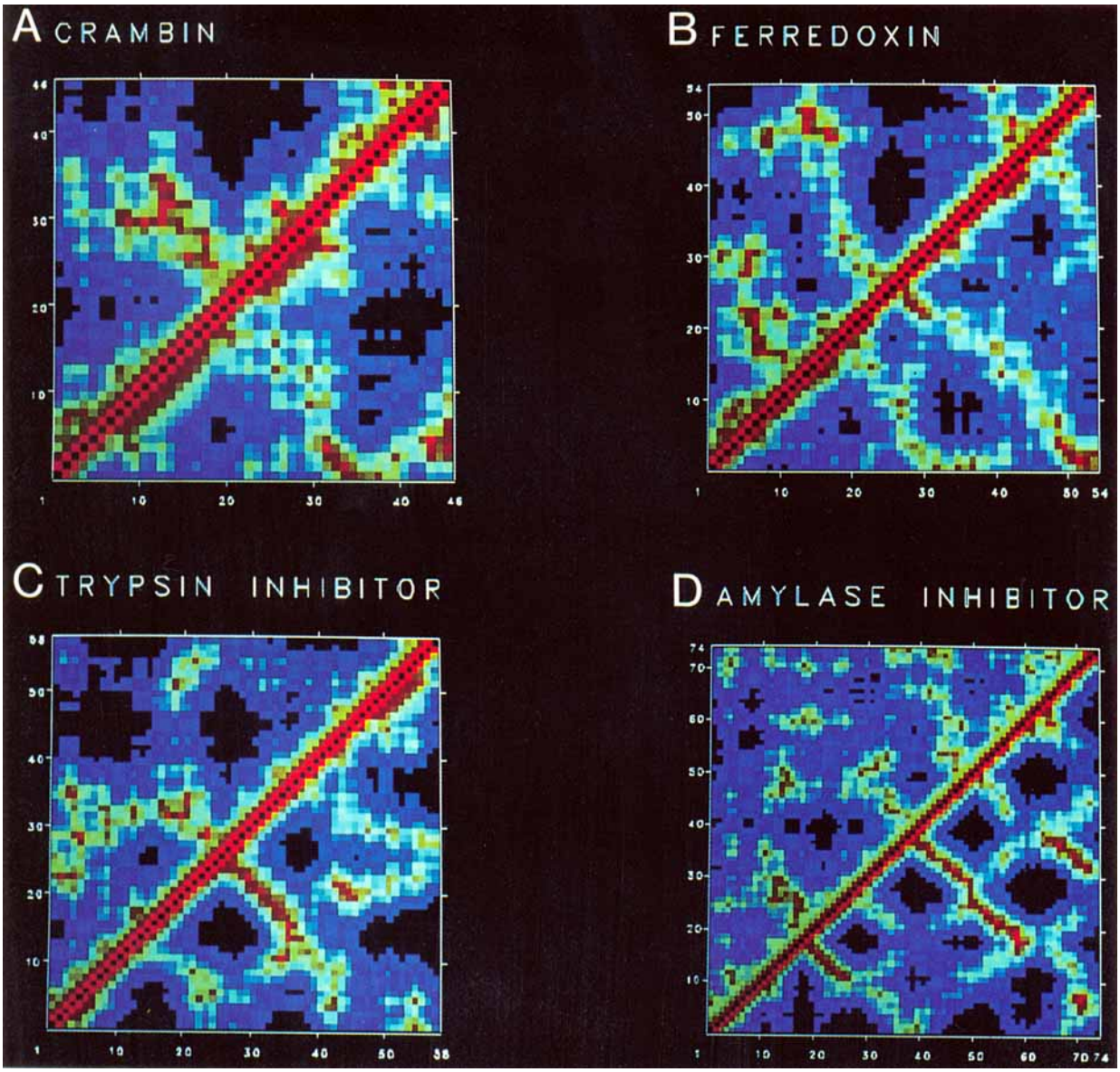


Fig. 7. α-carbon distance plots for four of the proteins studied. The native map is shown below and the final folded structure above the diagonal. Distances are colored spectrally with the closest distances appearing in red and the furthest in blue. All distances beyond 20 Å are shown in black. (**A**) 1CRN, (**B**) 1FDX, (**C**) 5PTI, (**D**) 1HOE.

long range contacts are also observed in the folded structures. The general off-diagonal patterns are quite similar, consistent with the topological agreement between backbones. The majority of close contacts (~60%) arise from spatially close α-carbons separated by less than 10 positions in sequence. Maps for the folded structures are less sharply defined than for their native proteins. Some of this fuzziness must result from not imposing the regularities of secondary structures observed in real proteins.[36] This is evidenced by the particularly poor correspondence between maps near the diagonal. Potential functions favoring secondary structure features commonly found in globular proteins might improve this situation. Such potentials would require a priori knowledge of locations where specific secondary structures exist in the protein's sequence. Incorporation of secondary structure potential functions into this procedure, or other local features such as those based on NMR data, is being investigated.

## DISCUSSION

Any theoretical determination of the folded state of a globular protein requires discrimination of the unique native conformation from the vast number

of possible conformations. This challenge is greatly reduced by confining conformational space to be native-like in shape[7] or by introducing *molecule-specific* biases in energy potentials.[9,11,16,48,49] Relaxation of these constraints is often balanced by model simplification. Typically there is a tradeoff between model precision and completeness of conformational exploration.

The simplified α-carbon model described here, using only amino acid sequence as a starting condition, appears useful for defining gross features of folded architecture. The folded forms generated compare favorably with native globular proteins with respect to residue energy per nonbonded contact (2–3 RT/contact), segment density ($\sigma = 0.65$) and location of surface loops and disulfide pairs. It can be argued that a portion of the rms error in these crude conformers is due to the lack of specification of secondary structure. The cubic lattice used does not precisely represent the details of α-helices and β-sheets. Currently available NMR data could define secondary structures and be used as constraints in further simulations starting from these folded structures.

The present analysis demonstrates that an extended linear chain composed of α-carbon elements restricted to lie on a simple cubic lattice can be folded into a native-like form. Chain energies evaluated using simple residue–residue potentials with preferences for neighboring amino acid types and crude consideration of local packing densities, when combined with a modified Monte Carlo sampling scheme, appear to be sufficiently discriminatory to define native-like chain topologies. Further testing and improvements to this methodology will be necessary to better assess its utility for understanding folded proteins.

## ACNOWLEDGMENTS

We thank Robert L. Jernigan, B. K. Lee, Jacob Mazur, Jacob Maizel, Danielle Konings, and Hugo Martinez for critical discussions on the topic of folded proteins. Generous technical support has been provided by George N. McGregor, Michael Scott, Jeffrey L. Garlough, William Boyer, Wayne Main, and Kai-Li Ting.

We thank the staff of the Biomedical Supercomputing Center, FCRDC, Frederick, MD for their assistance and access to the Cray X-MP supercomputer. Research sponsored, at least in part, by the National Cancer Institute, DHHS, under Contract N01-CO-74102 with Program Resources, Inc. The contents of this publication do not necessarily reflect the views or policies of the DHHS, nor does mention of trade names, commercial products, or organizations imply endorsement by the U. S. Government.

## REFERENCES

1. Kauzmann, W. Some factors in the interpretation of protein denaturation. Adv. Prot. Chem. 14:1–63, 1959.
2. Tanford, C. "The Hydrophobic Effect," 2nd ed. New York: John Wiley & Sons, Inc., 1979.
3. Dill, K.A. Dominant forces in protein folding. Biochemistry 29(31):7133–7155, 1990.
4. Crippen, G.M. Global optimization and polypeptide conformation. J. Comp. Physiol. 18(2):224–231, 1975.
5. Levitt, M., Sharon, R. Accurate simulation of protein dynamics in solution. Proc. Natl. Acad. Sci. U.S.A. 85:7557–7561, 1988.
6. Levinthal, C. Molecular model-building by computer. Sci. Am. 214(6):42–52, 1966.
7. Covell, D.G., Jernigan, R.L. Conformations of folded proteins in restricted spaces. Biochemistry 29(13):3287–3294, 1990.
8. Taketomi, H., Ueda, Y., Go, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. Int. J. Peptide Res. 7:445–449, 1975.
9. Kuntz, I.I., Crippen, G.M., Kollman, P.A., Kimelman, D. Calculation of protein tertiary structure. J. Mol. Biol. 106: 983–994, 1976.
10. Burgess, A.W., Scheraga, H.A. Assessment of some problems associated with the prediction of the three-dimensional structure of a protein from its amino-acid sequence. Proc. Natl. Acad. Sci. U.S.A. 72:1221–1225, 1975.
11. Levitt, M., Warshel, A. A computer simulation of protein folding. Nature (London) 253:694–698, 1975.
12. Skolnick, J., Kolinski, A., Yaris, R. Monte Carlo simulations of the folding of β-barrel globular proteins. Proc. Natl. Acad. Sci. U.S.A. 85:5057–5761, 1988.
13. Wilson, C., Doniach, S. A computer model to dynamically simulate protein folding with crambin. Proteins 6:193–209, 1989.
14. Ptitsyn, O.B., Rashin, A.A. A model for myoglobin self-organization. Biophys. Chem. 3:1–20, 1975.
15. Nemethy, G., Scheraga, H.A. Structure of water and hydrophobic bonding in proteins. II. model for the thermodynamic properties of aqueous solutions of hydrocarbons. J. Chem. Phys. 36:3401–3417, 1962.
16. Crippen, G.M., Snow, M.E. A 1.8 Å resolution potential function for protein folding. Biopolymers 29:1479–1489, 1990.
17. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
18. Richards, F.M. Areas, volumes, packing and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
19. Eisenberg, D.M., McLachlan, A.D. Solvation energy in protein folding and binding. Nature (London) 319:199–203, 1986.
20. Richards, F.M., Richmond, T. In: "Molecular Interactions and Activity in Proteins." Wolstenholme, G.E.M. (ed.), Ciba Foundation Symposium 60. Amsterdam: Excerpta Medica, 1978:23–45.
21. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–12, 1976.
22. Baumann, G., Frommel, C., Sander, C. Polarity as a criterion in protein design. Protein Eng. 2:329–334, 1989.
23. Meirovitch, H., Scheraga, H.A. Empirical studies of hydrophobicity. 2. Distribution of the hydrophobic, hydrophilic, neutral and ambivalent amino acids in the interior and exterior layers of native proteins. Macromolecules 13: 1406–1414, 1980.
24. Wertz, D.H., Scheraga, H.A. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. Macromolecules 11(1):9–14, 1978.
25. Creighton, T.E. "Proteins, Structures and Molecular Properties." San Francisco: W.H. Freeman, 1983.
26. Jaenicke, R. Folding and association of proteins. Prog. Biophys. Mol. Biol. 49 (1):117–237, 1987.
27. Karplus, M., Petsko, G.A. Molecular dynamics simulations in biology. Nature (London) 347:631–639, 1990.
28. McCrum, N.G., Read, B.E., Williams, G. "Anelastic and

Dielectric Effects in Polymeric Solids." New York: John Wiley, 1967.
29. Verdier, P.H., Stockmayer, W.H. Monte Carlo calculations on the dynamics of polymers in dilute solution. J. Chem. Phys. 36:227–235, 1962.
30. Monnerie, L., Geny, F. Monte-Carlo simulations of Brownian movement in a macromolecular chain. I. description of model and simulation. J. Chem. Phys. 66:1691–1697, 1969.
31. Hilhorst, H.J., Deutch, J.M., Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume. J. Chem. Phys. 63:5153–5161, 1975.
32. Lax, M., Brender, C. Monte Carlo study of lattice polymer dynamics. J. Chem. Phys. 67:1785–1787, 1977.
33. Ueda, Y., Taketomi, H., Go, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A three dimensional lattice model of lysozyme. Biopolymers 17:1531–1548, 1978.
34. Madras, N., Sokal, A.D. The pivot algorithm: A highly efficient Monte Carlo method for the self-avoiding walk. J. Stat. Phys. 50:109–186, 1988.
35. Kremer, K., Baumgartner, A., Binder, K. Monte Carlo renormalization of hard sphere polymer chains in two to five dimensions. Z. Phys. B 40(4):335–341, 1981.
36. Ramachandran, G.N., Sasisekharan, V. Conformation of polypeptides and proteins. Adv. Prot. Chem. 23:283–437, 1968.
37. 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.
38. Miyazawa, S., Jernigan, R.L. Equilibrium folding and unfolding pathways for a model protein. Biopolymers 21: 1333–1363, 1982.
39. McGregor, M.J., Cohen, F.E., Analysis of conformational tendencies in proteins. Current Opinion in Structural Biology 1(3):345–349, 1991.
40. Bryngelson, J.D., Wolynes, P.G., A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 30:177–188, 1990.
41. Tanaka, S., Scheraga, H.A. Medium and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9(6):945–950, 1976.
42. 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.
43. Sanchez, I.C. Phase transition behavior of the isolated polymer chain. Macromolecules 12:980–988, 1979.
44. Meirovitch, H. The scanning method with a mean-field parameter: Computer simulation study of critical exponents of self-avoiding walks on a square lattice. Macromolecules 18:563–569, 1985.
45. Correa, P. The building of protein structures from α-carbon coordinates. Proteins 7(4):366–377, 1990.
46. Phillips, D.C. Developments of crystallographic enzymology. Biochem. Soc. Symp. 31:11–28, 1970.
47. Nishikawa, K., Ooi, T., Ysogai, Y., Saito, N. Tertiary structure of proteins. I. Representation and computation of the conformations. J. Phys. Soc. Jpn. 32:1331–1337, 1972.
48. Levitt, M., Warshel, A. Folding and stability of helical proteins: carp myogen. J. Mol. Biol. 106:421–437, 1976.
49. Skolnick, J., Kolinski, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.
50. Dill, K.A. Theory for the folding and stability of globular proteins. Biochemistry 24:1501–1509, 1989.
51. Shakhnovitch, E.I., Finkelstein, A.V. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. Biopolymers 28:1667–1680, 1989.
52. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular protein structures. Proteins 6:87–103, 1989.
53. Flory, P.J., Fisk, S. Effect of volume exclusion on the dimensions of polymer chains. J. Chem. Phys. 44:2243–2248, 1966.
54. Go, N., Taketomi, H. Studies on protein folding, unfolding and fluctuations by computer simulation. III. Effect of short-range interactions. Int. J. Peptide Res. 13:235–252, 1979.
55. Creighton, T.E. Experimental studies of protein folding and unfolding. Prog. Biophys. Mol. Biol. 33:231–297, 1978.
56. Kim, P.S., Baldwin, R.L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. 51:459–489, 1982.
57. Bernstein, F.C., Koetzle, G.J.B., Williams, E.F., Meyer, M.D., Brice, J.R., Rodgers, O., Kennard, T., Shimanouchi, M., Tasumi, M. The Protein Data Bank; a computer-based archival file for macromolecular structure. J. Mol. Biol. 112:535–542, 1977.