Conformations of Folded Proteins in Restricted Spaces[†]

David G. Covell[*,‡] and Robert L. Jernigan[§]

Frederick Cancer Research Facility, Program Resources, Inc., Building 430, Frederick, Maryland 21701, and Laboratory of Mathematical Biology, National Cancer Institute, National Institutes of Health, Building 10, Room 4B-56, Bethesda, Maryland 20892

Received October 12, 1989; Revised Manuscript Received December 11, 1989

ABSTRACT: A new method is presented to examine the complete range of folded topologies accessible in the compact state of globular proteins. The procedure is to generate *all* conformations, with volume exclusion, upon a lattice in a space restricted to the individual protein's known compact conformational space. Using one lattice point per residue, we find 10^2–10^4 possible compact conformations for the five small globular proteins studied. Subsequently, these conformations are evaluated in terms of residue-specific, pairwise contact energies that favor nonbonded, hydrophobic interactions. Native structures for the five proteins are always found within the best 2% of all conformers generated. This novel method is simple and general and can be used to determine a small group of most favorable overall arrangements for the folding of specific amino acid sequences within a restricted space.

Most globular proteins consist of a polypeptide chain folded into a precise, highly organized, densely packed, three-dimensional structure. Understanding the molecular forces that determine these folded states remains one of the more compelling unsolved problems in the field of computational biochemistry. The demonstration of reversible denaturation for some proteins suggests that their three-dimensional structures are determined substantially by their amino acid sequences (Anfinsen, 1973). In fact, it has even been suggested that it is unlikely for a protein to adopt nonnative conformations under normal conditions (Flory, 1967).

The broad conformational space available to proteins appears to make the characterization of native or near native conformations difficult, both experimentally and theoretically. Folded states have been studied with molecular dynamics (Karplus & McCammon, 1983; Karplus, 1984), Monte Carlo generations (Kolinski et al., 1986), and other explorations of conformational space. Until now, computational limitations have precluded more thorough conformational explorations by these methods even for small proteins. Here we are taking a different approach and limiting our investigation to all feasible conformations of globular proteins within spaces defined by their native folded states. Specifically, we take a set of lattice points, each corresponding to the position of one amino acid in the protein's crystal structure, and generate all possible chain conformations upon these points. By means of this simple restriction, it then becomes possible to produce a complete set of conformations for small proteins that have substantial diversity. Furthermore, their stabilities can be evaluated on the basis of the statistics of close nonbonded amino acid pairs. For the five proteins studied, it appears that a small group of energetically favored conformations can be found that always includes the native conformations.

Lattice Geometry

Simple, low-resolution lattice or lattice-like models have often been used to represent diverse conformations (Orr, 1947; Flory, 1969; deGennes, 1979). Such models typically can account for chain flexibility and excluded volume effects, with some details of local and global chain packing. The present approach extends the lattice concept that defines and divides space by taking advantage of the reduction in the number of states afforded with an ideal lattice, together with advances in computational speed, to completely enumerate over the conformational space of small proteins.

α-Carbon coordinates are used to define the position of each amino acid, and the connections between adjacent positions define the system of virtual bond lengths, bond angles, and torsion angles. The proposed calculations require that the lattice model captures features of the spatial relationships between individual residues, but not at the price of gross distortions of virtual bond angles and virtual torsion angles. Levitt (1976) has shown that for globular proteins the strongest peak in the distribution of virtual bond angles is around 90°, with a weaker peak at 120°. Virtual torsion angle distributions display preferences for 45° and 210°. Furthermore, there is a simple correlation of bond angles and torsion angles: torsion angles at 45° exhibit bond angles in the neighborhood of 90° and torsion angles around 210° have bond angles near 120°.

Any regular lattice representation of a protein backbone uses a fixed set of virtual bond lengths, bond angles, and torsion angles. One choice for modeling globular proteins is the simple cubic lattice (SC) with the cube edge fixed at the virtual C^{α}–C^{α} spacing of 3.8 Å. However, its virtual bond angles of 90° and 180° and torsion angles of 0°, 90°, 180°, and 270° do not correspond closely to the observed preferences. The body-centered cubic lattice (BCC) provides four additional bond angles and eight additional torsion angles, if two bond lengths are included that permit connections along the edge of the cube (3.8 Å in length) as well as between the center of the cube and any corner (3.3 Å in length). Within these additional angle choices, the bond angle of 125.3° and the torsion angle of 45° are useful for improving fits. The face-centered cubic lattice (FCC) provides 4 additional options for bond angles and 16 additional options for torsion angles as compared to the simple cubic lattice, with bonds along the edge of the cube (3.8 Å in length) and between any corner and the center of a face (2.68 or 4.65 Å in length). Contrary to the BCC lattice, however, the bond angle choices for the FCC lattice can *exactly* match both preferred bond angles and the smaller of the two preferred torsion angles. The larger pre-

[†] Research sponsored, 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.

[‡] Frederick Cancer Research Facility.

[§] National Cancer Institute.

Table I: Comparison of Observed Virtual Bond Preferences and Nearest Lattice Values

	protein[a]	SC	BCC	FCC
bond angles	90	90	90	90
	120	90	125	120
torsion angles	45	0, 90	45	45
	210	180	225	215

[a] Observed protein preferences are taken from Levitt (1976).

Table II: Lattice Fits of C$^\alpha$ Coordinates

		RMS deviation (Å)		
protein[a]	no. of residues	SC	BCC	FCC
avian pancreatic polypeptide (1PPT)[b]	36	2.36	1.50	0.96
crambin (1CRN)[c]	46	2.70	1.46	0.96
rubredoxin (1RXN)[d]	52	2.51	1.89	1.04
ferredoxin (1FDX)[e]	54	2.25	1.65	1.07
neurotoxin (1NXB)[f]	62	3.60	2.32	1.07

[a] The five proteins have their X-ray crystal structures resolved to better than 1.5 Å (Bernstein et al., 1977). [b] Blundell et al., 1981. [c] Teeter, 1984. [d] Adman et al., 1977. [e] Adman et al., 1976. [f] Tsernoglou & Petsko, 1977.

ferred torsion angle of 210° is also quite close to the torsion angle of 215.3° available to the FCC lattice. Comparisons of the choices for each of these lattice types are given in Table I. It is clear that on this basis the FCC lattice is better than the BCC lattice, which in turn is better than the SC lattice for matching the preferred angles. The results of the least-squares fits to specific protein structures confirm this order (see Table II). The fit for the FCC lattice is best because the angles are well approximated, even at the price of added variability in bond lengths. Also, notably, it may not be the larger number of choices that matters so much as obtaining choices near the observed ones.

The FCC lattice representation has been chosen for our analysis. All proteins could be fit with this lattice to within 1.0-Å RMS deviation from the X-ray crystal structure. A least-squares procedure is used to fit lattice model coordinates to C$^\alpha$ coordinates. A discrete search of rotation angles about the x, y, z axes, using successively smaller grid sizes for the angles of rotation, is used to map a complete lattice onto an orientation that minimizes the sum of nearest-neighbor distances between lattice points and C$^\alpha$ positions. The edge of the cube for all three types of cubic lattices is fixed at the C$^\alpha$–C$^\alpha$ distance of 3.8 Å. The sum of squared neighbor distances is accumulated from the amino terminus to the carboxyl terminus. Lattice points already selected as nearest neighbors are thus unavailable to other residues. This procedure ensures that the chosen lattice representations are self-avoiding.

Alternative lattices such as the tetrahedral lattice yield fits consistently higher in RMS deviation than any of the cubic lattices. Furthermore, selection of an alternative single-point representation for each amino acid, such as C$^\beta$, or the centroid of the side chain, yields poorer fits in comparison with any of the cubic lattice models of C$^\alpha$ coordinates.

Chain Generation

The set of FCC lattice points corresponding to the C$^\alpha$ coordinates for each protein serves to define the space upon which conformations are generated. Bonds can be placed between any two lattice points that are within a 4.7-Å-radius sphere. The scheme to generate all possible paths is (1) select a starting point for the chain, (2) grow this chain by choosing the next step from the currently available nearest neighbors on the lattice, and (3) repeat the second step until all lattice points are occupied only once, or until a dead end is reached. The generation stops when all possible combinations of steps have been considered. Then a new starting position is selected and the process is repeated until all starting positions have been considered.

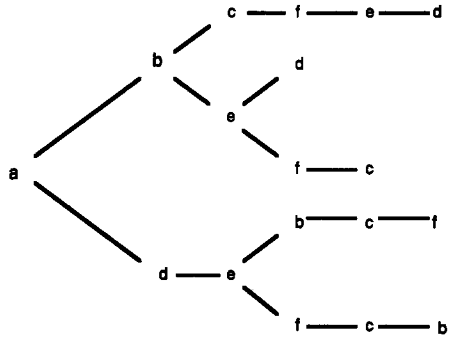
To illustrate this procedure, consider the 2 × 3 square lattice with points arbitrarily labeled a–f:

$$\begin{array}{ccc} d & e & f \\ \\ a & b & c \end{array}$$

Chain generation begins by constructing a list of all nearest-neighbor connections between two points. The matrix of all such neighbors is referred to in graph theory as an adjacency matrix (Christofides, 1975). This matrix contains elements equal to one for nearest neighbors and zero otherwise. The adjacency matrix for the present example is

$$\begin{array}{c|cccccc} & a & b & c & d & e & f \\ \hline a & 0 & 1 & 0 & 1 & 0 & 0 \\ b & 1 & 0 & 1 & 0 & 1 & 0 \\ c & 0 & 1 & 0 & 0 & 0 & 1 \\ d & 1 & 0 & 0 & 0 & 1 & 0 \\ e & 0 & 1 & 0 & 1 & 0 & 1 \\ f & 0 & 0 & 1 & 0 & 1 & 0 \end{array}$$

The process of completing a chain requires keeping track of the remaining choices available at each step of chain growth. This is simply a matter of updating the adjacency matrix by removal of points occupied by chain elements. For this example, the following paths would be taken to generate all chains starting from position a:



It is possible to complete only three unique chains from starting position a, namely, {a,b,c,f,e,d}, {a,d,e,b,c,f}, and {a,d,e,f,c,b}. The computer program to find these chains uses a depth first search strategy. This means that each path is continued to completion or until a dead end is reached. In the latter case, backward steps are taken until an alternative pathway can be found. The programming of such a recursive search strategy is straightforward.

Since chains starting at any of the four corners are rotationally equivalent, the only other unique starting position is b (or equivalently position e). Repeating the same procedure indicates that the single remaining unique chain on this lattice is {b,a,d,e,f,c}. Note that mirror images have not been considered in these cases. This example also serves to illustrate the sigificant reduction in the number of chain conformations occurring as a result of the restricted space and volume exclusion. On an unrestricted square lattice, an upper bound on the number of unique conformations of length 6 would be 2 × 3^3. Longer chains yield relatively larger reductions with these constraints. Some additional details of this method are discussed in Covell and Jernigan (1989).

Assessment of Conformations

The total collection of enumerated conformers for each protein is evaluated on the basis of pairwise, nonbonded contact energies. Nonbonded contacts, within 7.5 Å of each amino acid, are assumed to contribute to conformational stability. This assumption has its origin in hydrophobicity observations (Kauzmann, 1959; Eisenberg & McLachlan, 1986); i.e., buried hydrophobic residues have large numbers of hydrophobic neighbors. The sum E of all nonbonded contact pairs e_{ij} is an estimate of the energy of stabilization for each conformer.

$$E = \sum_{\text{contacts}} e_{ij}$$

Specification of e_{ij} according to residue–residue interactions is used to examine the role of sequence in protein stabilization. Details of the derivation of e_{ij} are given in Miyazawa and Jernigan (1985). Briefly, the contact energies, e_{ij}, used here are calculated as the net energy for forming a residue type i,j pair from residue type i, solvent and residue type j, solvent pairs. These values were derived from 42 high-resolution crystal structures by counting the numbers of residue–residue pairs ($N_{i,j}$), solvent–solvent pairs ($N_{\text{solvent,solvent}}$), and residue–solvent pairs ($N_{i,\text{solvent}}$,$N_{j,\text{solvent}}$) within a sphere of radius 6.5 Å centered at the average position of each amino acid side chain. These statistics were then used to define an equilibrium, from which contact energies were estimated as

$$\exp(-e_{ij}) = \frac{N_{i,j}N_{\text{solvent,solvent}}}{N_{i,\text{solvent}}N_{j,\text{solvent}}}$$

The strengths of these contact energies occur in decreasing order of their contribution to protein stability so that $e_{\text{hydrophobic–hydrophobic}}$ contributes the most, $e_{\text{hydrophilic–hydrophobic}}$ intermediate, and $e_{\text{hydrophilic–hydrophilic}}$ the least. The five proteins studied here were not included in the sample of 42 protein crystals used to obtain the e_{ij} values. The contact energies of Miyazawa and Jernigan (1985) did not include a repulsive component. Here, a constant term (+5.0 RT units) is added to each e_{ij} when the nonbonded contact distance is less than the sum of each residue's average radius [see Table III of Miyawaza and Jernigan (1985)].

Conformers for each protein were also evaluated on the basis of the hydrophobicity of residues in nonbonded contact. The total interaction energy H can be calculated as the sum of the hydrophobicities for each residue:

$$H = \sum_{\text{contacts}} h_{ij}$$

where

$$h_{ij} = h_i + h_j$$

Residue hydrophobicities were taken from the set of values in Miyazawa and Jernigan (1985) that were found to correlate well with the Nozaki–Tanford scale (Cornette, 1987).

The number of nonbonded contacts for each protein is constant for all conformations. This is because the lattice points defining the space in which folding must occur have a fixed number of neighbors. A connected path through these points removes a constant number of lattice points that could serve as nonbonded neighbors, regardless of conformation. Consequently, conformers cannot be distinguished on the basis of the number of nonbonded contacts.

Results

The primary focus is to examine the complete spectrum of backbone conformations for each protein in terms of their total nonbonded interaction energies, H and E. These results are summarized in Table III. Conformations have been ranked according to increasing E and H. Percentile rankings for the native structures are also given in this table. Native structures are found within the best 1.8% of all conformers generated for each protein when evaluated from contact energies (E). The best rankings are for crambin, ferredoxin, and sea snake neurotoxin, while the worst ranking is for avian pancreatic polypeptide. A 50-fold reduction in the number of favorable conformations is achieved from the complete set of conformers by using nonbonded contact energies. When conformers are evaluated on the basis of hydrophobicity (H), the ranking of the native structure was never better than that obtained with E (see Table III). In the worst cases, using H to discriminate finds the native structures of crambin and rubredoxin at the 8.8 and 12.3 percentiles, respectively.

Table III: Conformer Evaluations

protein	no. of conformers	ranking of native structure using[a]		no. of NB contacts[b]
		E	H	
1PPT	832	13 (1.8)	39 (4.7)	2.47
1CRN	15408	52 (0.3)	1355 (8.8)	3.30
3RXN	2258	31 (1.4)	277 (12.3)	3.00
1FDX	1952	4 (0.2)	11 (0.6)	2.96
1NXB[c]	3000	20 (0.7)	41 (1.4)	3.25

[a] Numbers are the numerical ranking of the native conformation and the numbers in parentheses are the percentile rankings based on energies (E) and hydrophobicities (H). [b] Average number of nonbonded contacts. [c] The enumeration of folded conformations from all lattice starting positions was not completed for 1NXB. Conformers were generated only for the starting position that corresponds to the amino-terminal C$^\alpha$ position. Generation of all conformations is not currently feasible for proteins of this size. For the five proteins studied, around 10000 chains could be generated per CPU hour on a Cray X-MP running CFT77 compiled Fortran code.

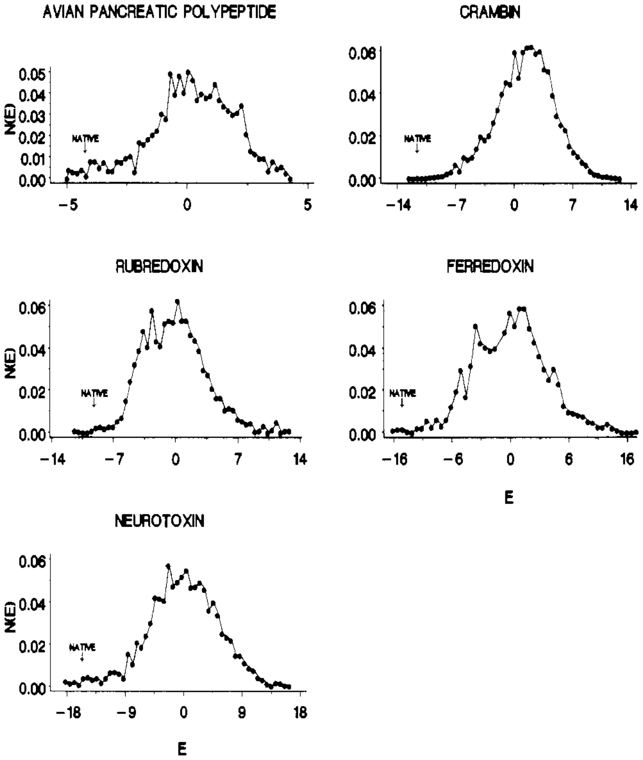


FIGURE 1: Distribution of energies for the five proteins studied: avian pancreatic polypeptide, crambin, rubredoxin, ferredoxin, and *Laticauda semifasciata* neurotoxin. Energies, E, are in dimensionless units of 10RT. The position of the native conformation is indicated by the labeled arrow.

Distributions of E for all conformers are shown in Figure 1. These distributions are generally smooth, are relatively symmetric, and cover a broad range of values of E. These

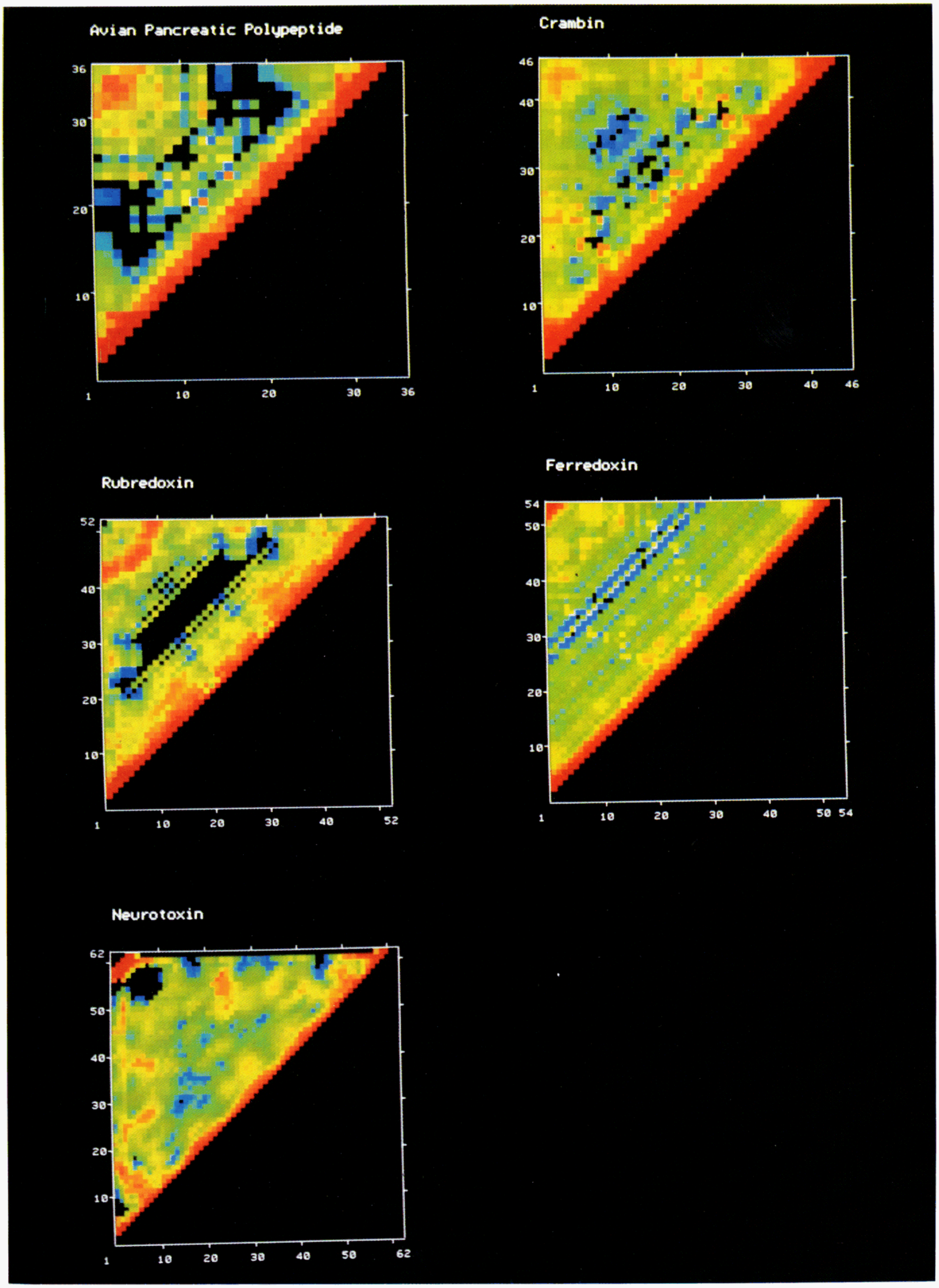


FIGURE 2: Distributions of nonbonded residue contacts for all conformers expressed as $-RT \ln (n_{ij}/N)$. Red (blue) areas signify the most (least) frequent nonbonded contacts. Totally black areas signify nonbonded contacts that *never* occur.

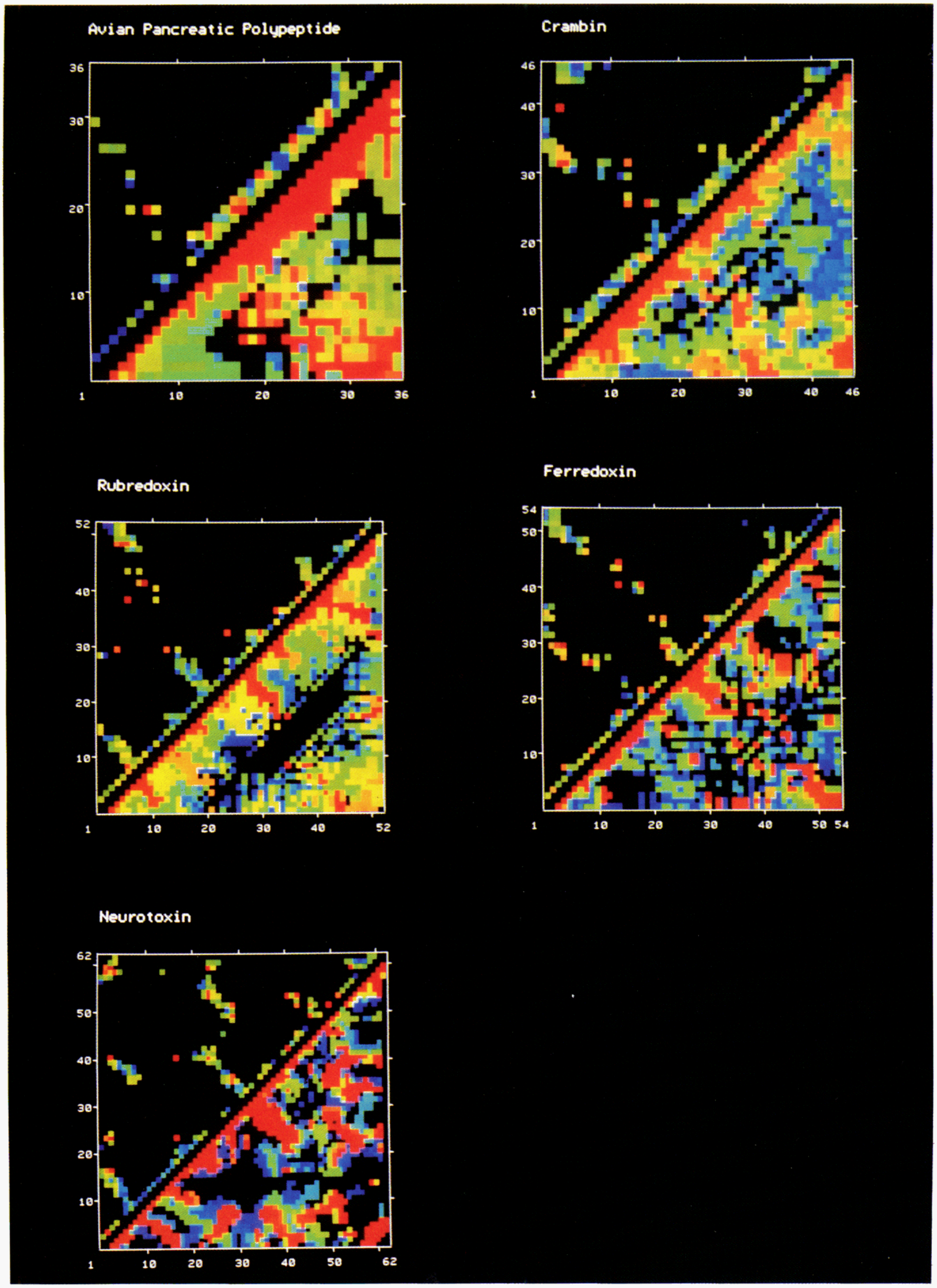


FIGURE 3: Weighted average energies for each nonbonded residue contact $\bar{e}_{ij}$ are plotted below the diagonal. Contacts that contribute the most (least) to lowering E are shown in red (blue). e_{ij} for the native protein is plotted above the diagonal. Colors for upper and lower regions are scaled separately from lowest (red) to highest (blue) energy.

features indicate that although the conformational space is quite restricted, the generated conformers yield a large variety of types of nonbonded contacts, which in turn lead to a continuum of values for E. The native conformations, indicated by the arrows in Figure 1, are always found in the favorable tail of the distribution.

The frequencies of nonbonded residue–residue interactions within the restricted space are summarized in Figure 2. These plots are essentially maps indicating the presence of a nonbonded contact between residue pairs. Such plots are often used to examine residue contacts within a single protein (Phillips, 1970; Nishikawa et al., 1972) and provide a useful picture of the three-dimensional details of the generated conformers. Notably, various types of secondary structures can be identified. Residues in an α-helical conformation appear as bands parallel and near the diagonal. An antiparallel arrangement of residues, either in an α-helix or β-sheet, appears as a band perpendicular to the diagonal; parallel arrangements appear as bands parallel but displaced from the diagonal.

Nonbonded contact frequencies for all conformers can be displayed at their respective positions on the contact map. These frequencies are calculated as the number of occurrences of nonbonded residues within 7.5 Å (n_{ij}) normalized by the total number of generated chains (N). These results describe how often a nonbonded residue contact occurs in the complete spectrum of conformations that were generated in the restricted space.

It is immediately clear from the contact plots that nearly all of the $n(n-1)/2$ residue–residue interactions do occur. Compact chains, such as those studied here, can be expected to have many intrachain contacts (Jernigan et al., 1988; Chan et al., 1989). The most frequent contacts are between "connected" neighbors, residues that are close to each other along the backbone chain. These interactions contribute to the stability of the protein by forming various types of secondary structures, primarily α-helices. The next most frequent contacts occur between ends of the chain, a result consistent with the observations of Thornton et al. (1983), showing that the amino and carboxyl termini of globular proteins are frequently found close to each other. Interactions between either end of the chain and the remainder of the chain are also common.

The five proteins studied here manifest those features described above, but with differences, as apparent in Figure 2. Note that, by definition, interactions between bonded pairs have been omitted here (i.e., along the diagonal). There are clear indications that specific types of secondary structures are preferred, depending on the protein. For example, a greater occurrence of α-helical structures exists for avian pancreatic polypeptide and crambin, proteins known to have substantial α-helical content. Conversely, rubredoxin, ferredoxin, and neurotoxin permit a greater occurrence of β-sheet structures, consistent with the primarily β-sheet composition of these proteins. Thus it appears that the overall shape of these compact proteins permits only a limited variety of secondary structure arrangements for packing backbone chain elements within this restricted space. But, the compactness leads to high fractions of secondary structure simply because regular structures can be packed more densely.

One region exists on these plots where contacts rarely occur. Nearly total exclusion of nonbonded contacts occurs in a region parallel to, but away from, the diagonal for some proteins. Contacts in this region represent a parallel arrangement of chain segments. One explanation for this excluded region is that parallel chains involve at least three segments, two forming the parallel strands and the third involved in a reversal. There are simply few possible arrangements of three chain segments in such a fashion when the space is restricted for these small proteins. Also the lengths of interacting antiparallel chain segments are restricted (i.e., secondary structures that produce bands perpendicular to the diagonal). This latter effect is probably due to the overall globular shape of these proteins and is related to the observation that the length of secondary structure elements in globular proteins is relatively short compared to their chain length (Chothia, 1984). These plots, however, clearly indicate that within each protein's restricted conformational space a nearly complete sampling of all possible nonbonded interactions is included.

The average energies for each nonbonded residue pair ($\bar{e}_{ij}$) are plotted below the diagonal in Figure 3. $\bar{e}_{ij}$ is calculated as

$$\bar{e}_{ij} = \frac{\sum_{\text{confs}} e_{ij} \exp(-E)}{\sum_{\text{confs}} \exp(-E)}$$

and represents the average contribution of nonbonded pairs to the total stability of each protein. In other words, each residue–residue contact plotted in Figure 2 is now plotted as the Boltzmann-weighted average. These plots indicate which interactions between residues favor hydrophobic pairs, i.e., frequently occurring pairs that contribute the most to a lower E. For comparison, each e_{ij} from the native structure is plotted above the diagonal in these figures. The clear pattern of energetically favored secondary structure elements in these plots indicates the importance of amino acid sequence in distinguishing between conformers. A strong α-helical pattern is evident at the amino terminus of avian pancreatic polypeptide and crambin while the most energetically favored regions for rubredoxin, ferredoxin, and neurotoxin occur for antiparallel β-sheets. There is also a strong correspondence between the most energetically favored nonbonded interactions in the total conformational space and the nonbonded C^α interactions in the native protein. Almost without exception, those nonbonded interactions that exist in the native folded state also appear as the most energetically favored interactions within the complete conformational space, i.e., when comparing any colored region in the upper half of the graphs with the most favored (red) regions in the lower half. This result indicates the potential of the proposed method for predicting which chain elements are most likely to produce favorable nonbonded interactions.

Plots of the relationship between E and the RMS deviation from the native conformer are shown in Figure 4. In general, the number of conformations and the range of values for E increase with increasing RMS. Groups of similar conformations with values of E lower than that of the native conformer do occur. These groups fall into two categories: those having small to intermediate RMS deviations and those having large RMS deviations. The first group, usually the smallest in number, frequently corresponds to small flips in the backbone conformation of the native structure that occur when bonds between two or four lattice positions are exchanged. This group includes avian pancreatic polypeptide and crambin. The second group occurs for crambin, rubredoxin, and ferredoxin and can be easily identified in Figure 4 as those points with lower than native E and large RMS values. These structures represent conformations that preserve energetically favored spatial arrangements of nonbonded pairs but do not have a nativelike backbone conformation. These conformations appear most frequently to arise from a reversal of the backbone

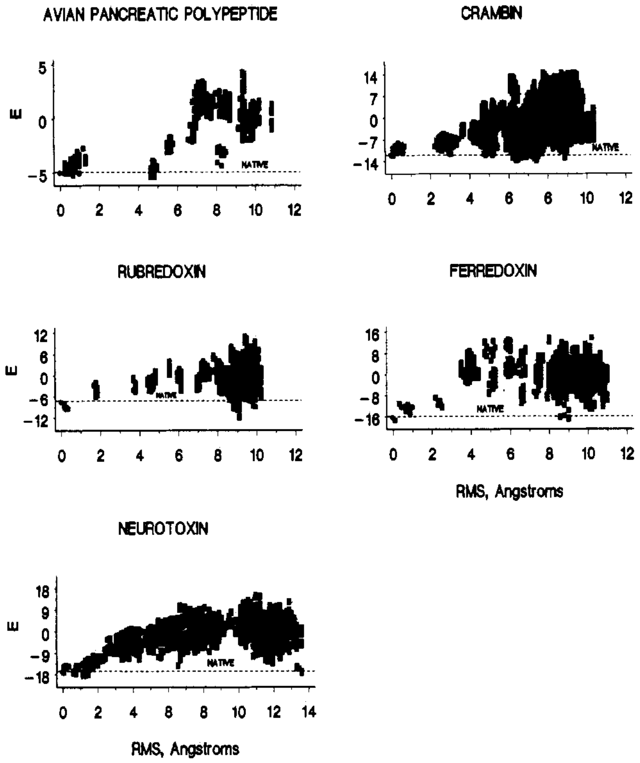
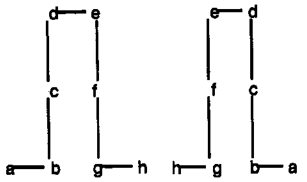


FIGURE 4: Calculated energies for all protein conformers as a function of RMS deviation from the native lattice conformation. Dark squares indicate the presence of at least one conformation. The position of the native lattice conformation is indicated by the filled circle at RMS = 0.0 and by the horizontal dashed line. Energies, E, are in dimensionless units of $10RT$.

direction (thus accounting for the large RMS value). In the present crude model at the level of one point per residue, identifying the direction of the chain can be a problem, both for fragments and for the entire chain. As an example, consider the following two loops:



In the present scheme both of these chain tracings would have the same nonbonded contact energy. Here the configuration of opposite sequence is equally likely.

Discussion

Our results indicate that for small proteins, at the level of one lattice point per amino acid, exhaustive enumeration of nativelike conformations is possible on a restricted lattice. The ensemble of conformations generated in this way produces chains with numerous intrachain contacts. However, a group of good conformers can be selected that favors the occurrence of hydrophobic–hydrophobic residue contacts over hydrophilic–hydrophobic and over hydrophilic–hydrophilic residue contacts (Figure 1). The native structure is always found within the best 1.8% of the total number of generated conformers (Table III). Accounting for these pairwise interactions, and simply evaluating each conformer in terms of neighboring hydrophobic and hydrophilic residues, does assist significantly in selecting the best conformers (Table III).

In the present study, conformers with lower than native contact energies occur because of statistical errors in e_{ij}, crudeness of the lattice, neglect of the heterogeneity of amino acid volumes, and too large a flexibility of the chain. Additional methods will be required to further narrow the choice among the lowest energy conformers. One approach would be to successively evaluate each of the good conformers obtained here by superimposing all atoms onto the lattice and refining these structures with conventional energy minimization (work in progress). Another approach would be to use a more complex lattice that explicitly includes side-chain positions.

The number of conformers generated is a function of the average number of lattice neighbors for bond selection. Lattice theories of chain molecules provide an estimate of the number of configurations as $(w/p)^z$, where z is the chain length, w is the conformational freedom or partition function per chain segment, and the value of p ranges (Dill, 1985; Flory, 1982) from 2.25 to e (=2.718). For the five proteins studied here, the average value of w/p is 1.18. Clearly, practical applications to long chains will require that w/p not greatly exceed 1. Pancreatic trypsin inhibitor, a 58-residue protein that we did not include in this report, is one example where the value of w/p is nearly equal to 1.0. Only 32 conformations were possible upon the lattice points selected to define the restricted space of folding. Half of these conformers originate from the lattice point that corresponds to the native amino terminus and the other half originate from the lattice point corresponding to the carboxyl terminus. The native conformation had the lowest contact energy, and no substantial differences were found between this protein and the five proteins studied in terms of goodness of lattice fit to the native C^α backbone and average number of nonbonded neighbors per lattice point. The number of conformers can also depend on precise details of the overall shape of each protein.

The utility of this approach for predicting folded protein conformations remains to be explored further. Its success depends on how well the appropriate restricted space can be delineated; such information may be available from preliminary X-ray crystallographic data (Kraut, 1958). We have not systematically pursued enumerations on less restricted lattices. However, conformations for avian pancreatic polypeptide and crambin were enumerated with additional freedom by including two more nonnative lattice points. The two additional lattice points were selected from the remaining FCC lattice points used to determine the lattice protein model. How to select the most appropriate additional lattice points is not clear, nor is any general relationship between additional lattice points and the number of conformers easily determined. In an attempt to determine an *upper* bound on the number of conformations, as well as to utilize the most buried points, the two lattice points selected had the largest number of lattice protein neighbors. With these additional points, the total number of conformers increases by an order of magnitude over those in Table III, but the number of conformers with lower than native energy is the same. Exhaustive enumerations on substantially less densely occupied lattices for proteins of the size considered here are not likely to be within immediate computational range. Advances in vector, parallel processing, and computer architecture should help overcome this limitation. Further strategies to impose deformations on the lattice may be fruitful.

One point of view of protein folding (Matouschek et al., 1989; Oas & Kim, 1989; Udgaonkar & Baldwin, 1988; Ptitsyn & Rashin, 1975; Richmond & Richards, 1978; Kim & Baldwin, 1982) is that peptides and proteins reach compact folded forms independently at early stages of folding. The present method could be useful for investigating these nonnative compact forms as well as conformational transitions. The present approach might be useful to crystallographers at early stages in their model development to indicate the likelihood of some alternative chain tracings. We have also begun to use the present approach to consider RNA folding in three dimensions. This latter application uses two lattice points per base, together with pairing and stacking energies (Tan et al., 1989). Binding of peptides to proteins is also being examined with this method (Jernigan et al., 1989). Selection of the appropriate lattice for enumeration of peptide conformations is based on knowledge of the three-dimensional structure of the target protein. The contact energy of each enumerated conformation of the peptide–protein complex can then be evaluated with the contact energies used here. Finally, our results lend strong support to the values of contact energies, e_{ij}, obtained from statistical analysis of the protein crystal database. These contact energies along with the known three-dimensional protein structure could be used to evaluate each residue's contribution to stability. This information could then be interpreted directly to indicate probable effects of site-directed mutagenesis (Reidhaar-Olsen & Sauer, 1988; Matsumura et al., 1988; Shortle & Meeker, 1989) on conformational stability.

Acknowledgments

We thank Boaz Shaanan, Wilma Olson, Michael Levitt, and Robert Baldwin for stimulating discussions. Generous technical support has been provided by Kai-Li Ting, William Boyer, Deborah Crum, and Michelle Spare. We are indebted to the staff of the Advanced Scientific Computing Laboratory, FCRF, Frederick, MD, for their assistance and for access to the Cray X-MP supercomputer. This access provided a substantial role in the motivation and development of this project.

References

Adman, E. T., Sieker, L. C., & Jensen, L. H. (1976) *J. Biol. Chem. 251*, 3801–3806.

Adman, E. T., Sieker, L. C., Jensen, L. H., Bruschi, M., & Le'Gal, J. (1977) *J. Mol. Biol. 112*, 113–120.

Anfinsen, C. B. (1973) *Science 181*, 223–230.

Bernstein, F. C., Koetzle, G. J. B., Williams, E. F., Meyer, M. D., Brice, J. R., Rodgers, O., Kennard, T., Shimanouchi, M., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P., & Wu, C. W. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4175–4179.

Chan, H. S., & Dill, K. A. (1989) *J. Chem. Phys. 90*, 492–509.

Chothia, C. (1984) *Annu. Rev. Biochem. 53*, 537–572.

Christofides, N. (1975) *Graph Theory, An Algorithmic Approach*, Academic Press, New York.

Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., & DeLisi, C. (1987) *J. Mol. Biol. 195*, 659–685.

Covell, D. G., & Jernigan, R. L. (1989) *Int. Conf. Supercomput., 4th 2*, 357–360.

deGennes, P. G. (1979) *Scaling Concepts in Polymer Physics*, Cornell University Press, Ithaca, NY.

Dill, K. A. (1985) *Biochemistry 24*, 1501–1509.

Eisenberg, D. M., & McLachlan, A. D. (1986) *Nature 319*, 199–203.

Flory, P. (1967) in *Conformations of Biopolymers* (Ramachandran, G. N., Ed.) Vol. 1, pp 339–363, Academic Press, New York.

Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience Publishers/Wiley and Sons, New York.

Flory, P. J. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 4510–4514.

Jernigan, R. L., Sarai, A., Mazur, J., & Covell, D. G. (1988) *Int. Conf. Supercomput., 3rd 1*, 197–200.

Jernigan, R. L., Margalit, H., & Covell, D. G. (1990) in *Theoretical Biochemistry and Molecular Biology, A Computational Approach*, Adenine Press, Schenectady, NY.

Karplus, M. (1984) *Adv. Biophys. 18*, 165–190.

Karplus, M., & McCammon, J. A. (1983) *Annu. Rev. Biochem. 52*, 263–300.

Kauzmann, W. (1959) *Adv. Protein Chem. 14*, 1–63.

Kim, P., & Baldwin, R. L. (1982) *Annu. Rev. Biochem. 51*, 459–489.

Kolinski, A., Skolnick, J., & Yaris, R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 7267–7271.

Kraut, J. (1958) *Biochim. Biophys. Acta 30*, 265–270.

Levitt, M. (1976) *J. Mol. Biol. 104*, 59–107.

Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–239.

Matouschek, A., Kellis, J. T., Jr., Serrano, L., & Fersht, A. R. (1989) *Nature 340*, 122–126.

Matsumura, M., Becktel, W. J., & Matthews, B. W. (1988) *Nature 334*, 406–410.

Miyazawa, S., & Jernigan, R. L. (1985) *Macromolecules 18*, 534–552.

Nishikawa, K., Ooi, T., Ysogai, Y., & Saito, N. (1972) *J. Phys. Soc. Jpn. 32*, 1331–1337.

Oas, T. G., & Kim, P. (1989) *Nature 336*, 42–48.

Orr, W. J. C. (1947) *Trans. Faraday Soc. 43*, 12–27.

Phillips, D. C. (1970) *Biochem. Soc. Symp. 31*, 11–28.

Ptitysyn, O. B., & Rashin, A. A. (1975) *Biophys. Chem. 3*, 1–20.

Reidhaar-Olson, J. F., & Sauer, R. T. (1988) *Science 241*, 53–57.

Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng. 6*, 151–176.

Richmond, T. J., & Richards, F. M. (1978) *J. Mol. Biol. 119*, 537–555.

Shortle, D., & Meeker, A. K. (1989) *Biochemistry 28*, 936–944.

Tan, R. K., Prabhakaran, M., Tung, C. S., & Harvey, S. C. (1988) *Comput. Appl. Biosci. 4*, 147–151.

Teeter, M. M. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 6014–6018.

Thornton, J. M., & Sibanda, B. L. (1983) *J. Mol. Biol. 167*, 443–460.

Tsernoglou, D., & Petsko, G. A. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 971–974.

Udgaonkar, J. B., & Baldwin, R. L. (1988) *Nature 335*, 694–699.