the SGAL's in terms of the relative positions of I and II, rather than of I and I + II and, for this reason, we refer to II as a loop in the following discussion.

(41) Klapper, M. H.; Klapper, I. Z. *Biophys. J.* **1980**, *32*, 216. *Biochim. Biophys. Acta* **1980**, *626*, 97.

(42) Krigbaum, W. R.; Komoriya, A. *Biochim. Biophys. Acta* **1979**, *576*, 204.

(43) Meirovitch, H.; Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1398.

(44) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1406.

(45) Meirovitch, H.; Scheraga, H. A. *Macromolecules* **1981**, *14*, 340.

(46) BPTI and HIPIP are two proteins with radial distributions of hydrophobic and hydrophilic amino acids that differ from the general behavior.[43,44] Therefore, the average distribution cannot be used in these cases.

(47) Minimization of the energy with respect to all 267 backbone and side-chain dihedral angles is very time-consuming. We found it more efficient to minimize in two cycles. In the first, the energy was minimized with respect to all backbone dihedral angles and all side-chain dihedral angles χ^1 and χ^2 (which have the largest effect on the orientation of the side chain), i.e., 196 variables. In the second cycle, the energy was minimized with respect to all 155 side-chain dihedral angles. In one trial, 227 dihedral angles were varied (ϕ, ψ, χ^1, χ^2, and χ^3), but in all other trials only 155 or 196 dihedral angles were varied.

(48) Dennis, J. E.; Mei, H. H. W. Technical Report No. 75-246, 1975, Department of Computer Science, Cornell University, Ithaca, N.Y.

(49) Pottle, C.; Pottle, M. S.; Tuttle, R. W.; Kinch, R. J.; Scheraga, H. A. *J. Comput. Chem.* **1980**, *1*, 46.

(50) Swenson, M. K.; Burgess, A. W.; Scheraga, H. A. In "Frontiers in Physicochemical Biology"; Pullman, B., Ed.; Academic Press: New York, 1978; p 115.

(51) Deisenhofer, J.; Steigemann, W. *Acta Crystallogr, Sect. B* **1975**, *31*, 238.

(52) We compare our calculated structures with NI rather than with NF because NI is closer than NF to the experimental X-ray structure. Nevertheless, the energy of NF serves as a reference for the lowest energy attainable by our procedure.

Differential Geometry and Polymer Conformation. 3. Single-Site and Nearest-Neighbor Distributions, and Nucleation of Protein Folding[1]

S. Rackovsky[2a] and H. A. Scheraga*[2b]

Biophysics Department, Weizmann Institute of Science, Rehovot, Israel, and
Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.

Received February 17, 1981

ABSTRACT: A differential geometric representation of polymer chains is used to study protein backbone structure on the four- and five-C^α length scales. The analysis of the distribution of four-C^α units in the curvature–torsion (κ, τ) plane reveals that there are islands of structure which occur with greater-than-random probability. These correspond to the well-known types of backbone structure—extended (E), right-handed α-helical (A_R), and flat-bend (A_0) structure. It is shown that there are three distinct types of extended four-C^α structure—left-handed twisted (E_L), right-handed twisted (E_R), and nearly flat (E_0). The E_0 and E_L structures form a structural continuum, but the E_R region is separated from this continuum by a region of low occupation. The A_R and A_0 regions also form a continuum. These high-frequency islands are distributed throughout the occupied region of the (κ, τ) plane and include structures of all types except that in the A_L (left-handed α-helical) region. The high frequency of occurrence of these structures suggests that they are of low energy and therefore likely to occur in nucleation structures in the folding of the denatured molecule. It follows from the fact that these high-frequency structures occur throughout the occupied region, however, that there is *relatively low selectivity in nucleation on the four-C^α scale*, with structures representative of the entire occupied region being potential nuclei on the four-C^α scale. Extension of the analysis to the five-C^α scale shows that the high-frequency structures on this scale are made up of combinations of the high-frequency four-C^α structures and that *a higher degree of selectivity in nucleation appears on the five-C^α length scale* than on the four-C^α scale. Substantial differences are noted in the frequency of occurrence of the various combinations. Within the extended region, for example, the most frequently occurring structures are E_0E_0, with E_RE_R being the least frequent. A study of the correlation between nearest-neighbor four-C^α structures reveals that certain *pairs* of four-C^α structures have a low probability of occurring. A_0A_R and A_RA_0 have a low positive correlation, while E_XA_R and A_RE_X (where X is R, L, or 0) have a negative correlation, indicating that these structures tend to *avoid* pairing. Five-C^α components of nucleating structures are likely to be those which both occur with high frequency and show positive correlation between their four-C^α components. Of these, the most frequently occurring are E_XE_X and A_RA_R, which are repeating (regular) structures. The A_0 four-C^α structure plays an important role in nucleation, despite its relative numerical infrequency, because it is the principal four-C^α structure which forms potential five-C^α nucleating structures which are nonregular. Larger nuclei, which are not considered explicitly in this paper, presumably arise as additional four-C^α units associated with five-C^α nuclei already present in the folding chain.

I. Introduction

One of the significant aspects of protein architecture is the presence, in molecules with widely differing function and amino acid sequence, of certain well-defined structural features. Historically, the first such features to be noted were the α helix and the extended strand (which associates with other strands to form the β sheet). These two structures are particularly easy to observe and classify because they are built up of repeating units (residues) with similar conformation. They are therefore capable of occurring, and of being observed, on a wide variety of backbone length scales, from a single residue to tens of residues.

It later became clear that there is at least one important structural feature which occurs on a single, well-defined length scale. This is the bend, or chain reversal, whose

0024-9297/81/2214-1259$01.25/0 © 1981 American Chemical Society

presence is determined entirely by the structure of a backbone segment consisting of four successive α carbons—or, equivalently, by the values of two successive sets of (ϕ, ψ) dihedral angles. Unlike α helices or extended strands, bends are not built up of residues with essentially *repeating* values of (ϕ, ψ). In fact, a rather wide variety of combinations of $(\phi_i, \psi_i; \phi_{i+1}, \psi_{i+1})$ form bends,[3,4] with quite similar C^α arrangements.[5] Hence, bends differ from α helices and extended structures (in β sheets) in the manner in which they are stabilized relative to their single-residue energies. α helices and extended structures are each stabilized by (various interactions, including) a single type of interresidue hydrogen bond. Bends, on the other hand, may be stabilized by hydrogen bonds in a number of ways[6] or involve no hydrogen bonds at all, depending on the particular values of (ϕ, ψ) which obtain. It is also possible for bends to be stabilized in still other ways—e.g., by hydrophobic interactions—completely external to the bend region itself.[7]

One then can raise the following question: Are there other well-defined structural features in proteins, as yet unobserved because they occur on length scales which are not easily accessible through conventional structure representations?

The identification and characterization of specific structural features of proteins is important because such structures are believed to be essential for the correct self-organization of the protein molecule, a process which is crucial to its proper function. For example, it has been proposed[8] that the nucleation regions which initiate folding are hydrophobic pockets that form around bends; another theory of folding[9] suggests that nucleation occurs by formation and disintegration of α helices.

Before proceeding, it is necessary to define the term nucleation, as we shall use it in this paper. The denatured protein may be regarded as a fluctuating "random coil". We suggest that, as renaturing conditions are created, ordered structures form in the fluctuating chain; their lifetimes are much longer than those of any conformations which form in neighboring parts of the chain. We regard these long-lived structures as nuclei. These structures need not all form simultaneously in the fluctuating coil, and, indeed, it may be necessary for them to form in a specific sequence to attain the correct globular folding. Furthermore, not all of these structures need individually cause a reduction in chain dimensions. Previous work[8] has been directed toward the identification of the most stable structure responsible for a decrease of chain dimensions. We shall be interested in the entire spectrum of possible nucleating structures.

Nucleating structures of the type that we have described can be of various lengths. Such structures must be built up residue-by-residue, beginning from very short lengths of backbone which assume the proper conformation solely under the influence of localized, short-range interactions. Thus, each nucleus undergoes an elongation process until it reaches its ultimate length. It then contributes to the folding process either by bringing about favorable interactions between other parts of the chain or by participating in such interactions itself.

In this paper, we use a differential geometric (DG) representation of polypeptide chains to extend the search for significant structures systematically to longer length scales than that treated by the (ϕ, ψ) representation. As was demonstrated in papers 1[10] and 2[5] of this series, the DG representation operates on a four-C^α length scale (Figure 1a). It is therefore able, for example, to give a structurally clear description of bends, which is not easily provided by the (ϕ, ψ) representation.[5] We shall begin by investigating the distribution of values of curvature and torsion (κ_i, τ_i) of four-C^α units (defined in paper 1[10]), with a view to characterizing structure on this scale in a systematic fashion. We shall then investigate the distribution of nearest-neighbor $(\kappa_i, \tau_i; \kappa_{i+1}, \tau_{i+1})$ pairs. This distribution provides information about the structure of five-C^α units (Figure 1b) and represents a first step in the extension of the characterization of significant structures to still longer length scales.

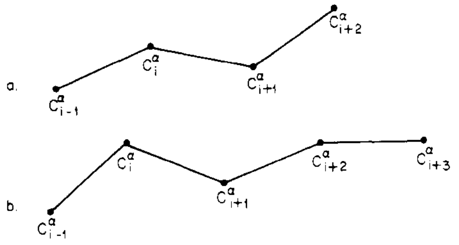

Figure 1. (a) Four-C^α unit whose conformation is described by (κ_i, τ_i). (b) Five-C^α unit whose conformation is described by $(\kappa_i, \tau_i; \kappa_{i+1}, \tau_{i+1})$. These are representations of virtual-bond structures, and the quantities κ_i and τ_i were defined in the first paper.[10]

We shall adopt the hypothesis that those structures that occur with high frequency in native proteins are of low energy and therefore have a high probability of forming in the denatured chain under renaturing conditions. We shall identify these as possible nucleation structures. This is not to suggest that four- or five-C^α structures are the *complete* nucleating structures. This is possibly, but not necessarily, true. Examination of protein structures on a variety of length scales should reveal that some high-frequency structures have well-defined length scales and others do not—precisely the situation that we noted above in connection with bends, helices, and extended strands. We *do* suggest, however, that the four- and five-C^α nucleating structures that we shall identify form parts of complete nucleating structures and therefore are potentially important in the process of growth of nuclei. Their formation constitutes an early stage in the growth of the individual nuclei. Thus, our goals in this paper are (1) to delineate the structural types which are significant on the four-C^α and five-C^α length scales; (2) to provide a quantitative picture of the relative importance of these different structural types in native proteins; and (3) to draw conclusions from this analysis about the nucleation process which leads to correct folding of the protein.

II. Single-Site Distribution

In paper 1,[10] we explored the localization of various ordered backbone structures in (κ, τ) space by determining (κ_i, τ_i) at residues which had been determined by (ϕ, ψ) criteria to be in the conformational state of interest. We now adopt a more general approach and ask what types of characteristic structures occur on the single-site (four-C^α) length scale in (κ, τ) space, without reference to the values of (ϕ, ψ). [We shall use the term "single site" throughout this paper to denote an individual (κ_i, τ_i) pair, ascribed by convention[10] to residue i. It should be borne in mind, however, that (κ_i, τ_i) is determined by the locations of the *four* atoms $C^\alpha{}_{i-1}$ to $C^\alpha{}_{i+2}$ (Figure 1a) or alternatively by *two* pairs of dihedral angles $(\phi_i, \psi_i; \phi_{i+1}, \psi_{i+1})$.]

This more general approach is particularly important because of the degeneracy in terms of the values of (ϕ, ψ) which is observed on the four-C^α length scale.[5] This degeneracy implies that very similar four-C^α virtual-bond structures, having widely differing combinations of $(\phi_i, \psi_i; \phi_{i+1}, \psi_{i+1})$, can exist. It follows that an analysis in terms

Table I
Protein Sample

1. bovine pancreatic trypsin inhibitor
2. concanavalin A
3. carboxypeptidase A
4. clostridial flavodoxin
5. thermolysin
6. oxidized chromatium high-potential iron protein
7. staphylococcal nuclease
8. sea lamprey hemoglobin
9. sperm whale myoglobin
10. hen egg white lysozyme
11. ribonuclease S
12. subtilisin BPN' (Novo)
13. rubredoxin
14. carbonic anhydrase
15. tuna cytochrome *c* (oxidized inner)
16. D-glyceraldehyde 3-phosphate dehydrogenase
17. papain
18. carp myogen
19. human plasma prealbumin
20. ferrodoxin
21. α-chymotrypsin
22. chicken triose phosphate isomerase

of (ϕ, ψ) may well fail to reveal certain structural types which, though well-defined on the (κ, τ) length scale, are diffuse in terms of the values of (ϕ, ψ). An obvious example of such a structural type is the bend, which was discussed in the Introduction. The traditional (ϕ, ψ) classification,[3,4] which produces 11 bend types, does not easily reveal the fact that bends form a continuum of structural types, ranging from segments of α_R-helical structures through flat bends to segments of α_L-helical structures.[5] There is every reason to suppose that the same effect holds in other regions of (κ, τ) space.

We adopt a graphical approach to the problem and begin by constructing the observed distribution of values of (κ, τ) in a sample of proteins from their X-ray coordinates. A preliminary form of this distribution was exhibited in Figure 3 of paper 1.[10] For the current study, the sample of 13 proteins used in paper 1 was expanded by the addition of nine more proteins, so chosen that, in the combined sample (as in the original sample of paper 1), there are few homologies and the coordinates available are of reasonable quality. The proteins used are listed in Table I; all coordinates were obtained from the Brookhaven Protein Data Bank.

The resulting distribution is shown in Figure 2, plotted as the number of four-C$^\alpha$ segments per (0.1 × 0.1) rad/Å square in the (κ, τ) plane, independent of residue type. This is essentially a top view of a three-dimensional bar graph, in which the heights of the bars are given by the occupancy numbers of the squares. The essential identity of the occupied region of Figure 2 with that of the smaller sample is readily seen in Figure 2, in which the occupied squares of the graph have been assigned an arbitrary numbering for convenient reference (small numbers in lower right hand corners of squares). Squares 1–66 are occupied in the original 13-protein sample of paper 1. Squares 70–73 and 76 are occupied only in the 22-protein sample of the current work. (The omission of some numbers from the arbitrary numbering of squares arose from an earlier version of this work and is of no significance.) There are only seven points in the five new squares. Comparison of the two distributions shows that there is also no significant change in the relative proportion of points falling in the various regions of the (κ, τ) plane, in passing from the original sample of 2127 points to the current sample of 3754 points. We therefore believe that the two samples both represent reasonably well the actual distribution of conformational preferences in proteins.

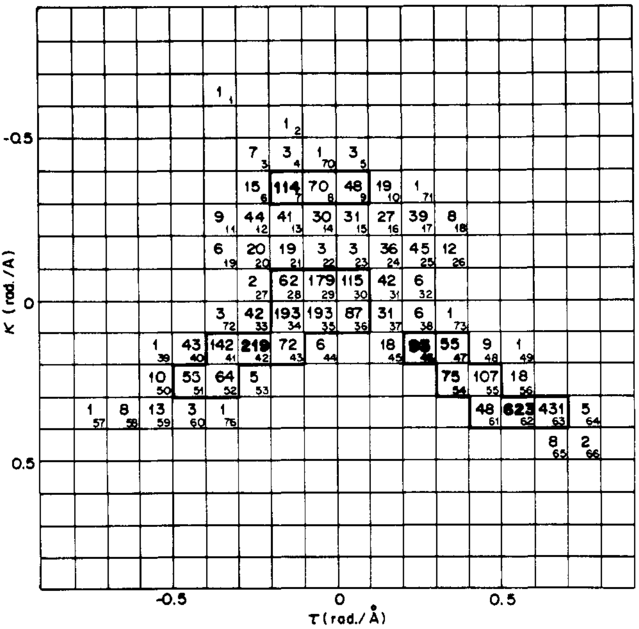

Figure 2. Untransformed single-site distribution of values of (κ, τ) for the sample of 22 proteins listed in Table I. The islands of greater-than-random occupation are indicated by heavy outlines. The large number in each (0.1 × 0.1) rad/Å square is the number of points observed to fall in that square, and the small number is an arbitrarily assigned numbering of the squares. Occupation numbers in boldface indicate that those squares are maxima in the distribution. Shaded area (squares 46, 47, and 54) indicates one of the two peaks on an island with two maxima (see text and Appendix II). The squares comprising the various regions of peaks in this distribution (A_0, E_L, etc.) are identified in Table II.

Before proceeding, it is necessary to recall briefly a number of facts about the (κ, τ) plane which were developed in paper 2[5] (to which the reader is referred for mathematical details). Two in particular will be of interest.

1. There is a discontinuity in κ and τ as the conformation of a four-C$^\alpha$ segment is changed continuously, at a conformation where the factor p_ip_{i+1} (which enters into the formulas for κ_i and τ_i) changes sign (p_ip_{i+1} assumes the values ±1). The quantity p_i is defined in eq 6 of paper 1[10] as the scalar product of two nonperpendicular unit vectors at the ith and (i − 1)th α-carbon atoms, respectively; it is a factor which guarantees that the unit normal in the local reference frame at residue i has a positive projection on the unit normal in the local reference frame at residue i − 1. As a result, there are certain regions of the (κ, τ) plane where structures that are conformationally very similar appear as widely separated points. In paper 2, it was shown that the structures affected by this discontinuity are bends and allied structures. It was also shown that a transformation exists (eq 16 of paper 2) which brings structures for which $p_ip_{i+1} = +1$ (which are in the minority) to new points in the (κ, τ) plane which are adjacent to those representing conformationally similar structures for which $p_ip_{i+1} = -1$. In Appendix I, we apply this transformation to the distribution of Figure 2 and perform an analysis of the resulting distribution which demonstrates that the conclusions that we shall draw from Figure 2 are correct. This is important because we shall use the untransformed values of (κ, τ) to construct the nearest-neighbor distribution of section III. We do this because the untransformed distribution provides a natural method for distinguishing between the right- and left-handed twisted (A_R and A_L) bends, on the one hand, and the structurally adjacent flat (A_0) bends, on the other.[11] We shall show that

these structures have distinctly different behavior in the nearest-neighbor distribution.

2. There is also a discontinuity in κ for four-C$^\alpha$ structures with $p_ip_{i+1} = -1$ at the conformational point where κ changes sign. It was demonstrated in paper 2 that this leads to a butterfly-shaped gap centered about the τ axis.

Both of these factors will be taken into account (although not explicitly) in the analysis of the distributions.

We now return to the analysis of Figure 2. If the sample of 3754 points were distributed uniformly over the 71 squares which constitute the occupied region of Figure 2, as might be if each four-C$^\alpha$ unit were equally likely to assume any conformation in the occupied region, the population per square would be 52.9.[12] We therefore ask which squares are occupied with greater-than-random frequency. These squares form three separate islands, which are indicated in Figure 2 by heavy outlines. [To account for possible statistical error, we included squares with population ≥48 (=0.9 × 53), where the factor 0.9 was chosen arbitrarily.] It should be noted that, whereas the squares included in these islands contain only 30% of the available area, they contain 81.1% of the residues—indicating the utility of the island approach for identifying high-frequency structures.

What types of four-C$^\alpha$ structure are represented by the points contained in the islands of high frequency? The answer to this question becomes clear from a comparison of Figure 2 with Figures 6–8 of paper 1. The island embracing squares 7–9 (which we shall refer to as island I) corresponds precisely to the low-torsion bend region of Figure 8, paper 1. The island which includes the origin (which we shall refer to as island II) substantially overlaps the distribution of extended structures of Figure 7, paper 1. The remaining island (island III) includes the principal α_R-helical region of Figure 6, paper 1, and the (coincident) α_R-helical bend region of Figure 8. It also includes part of the region identified as including extended structures in Figure 7, paper 1.

A more detailed analysis of the individual islands is in order. We begin with island II, which includes the largest part of the extended region. There is a striking breadth of structural types included in this island. These range from structures which are very nearly planar zigzags (such as those concentrated near the origin, in squares 29, 30, 35, and 36) to those which are nearly α_L-helical (square 51). There is a very strong predominance of left-handed structures.[13–19] The population maximum at square 42 falls in a region which is both bent (in terms of change of chain direction, as shown by κ) and twisted (in a left-handed sense, as shown by τ) relative to a perfect planar zigzag form.

There is only one maximum in island II (that at square 42), and this island clearly contains a continuum of structural types. Nevertheless, it is useful and meaningful to distinguish between the nearly planar structures centered about the origin and the more twisted, more bent structures clustered about the maximum. We shall denote these as E$_0$ and E$_L$ structures, respectively. The subscript in E$_0$ is intended to denote the fact that these structures are almost flat; of course, since they are usually not perfectly planar, they do have a handedness. The E$_L$ structures have a pronounced left-handed twist.[13–19] We use a capital E to emphasize the fact that the structures defined by these values of (κ, τ) are on a longer scale than those defined by a (ϕ, ψ) pair which falls in the extended (ϵ) region of the (ϕ, ψ) plane. To make our notation precise, we define as E$_0$ those structures which are represented by points which fall in squares 28–30 and 34–36. E$_L$ structures are those which fall into squares 41–43 and 52. Because the main body of the A$_L$ region (squares 58–60) is sparsely occupied, square 51 is assigned to the E$_L$ region, although the symmetric right-handed square (55) is assigned as an outlier of the A$_R$ region. This flexibility of assignment illustrates strikingly the general fact that there is a *continuous gradation* of four-C$^\alpha$ structural types, rather than a collection of discrete varieties.

An obvious question is whether there is a corresponding E$_R$ structure in the region of high-frequency structures. One would expect such structures to have the same values of κ as E$_L$ structures but values of τ opposite in sign. Inspection reveals that the maximum at square 46 in island III and the neighboring squares of that island (i.e., squares 47 and 54) indeed constitute the E$_R$ region. It should be noted that the populations of squares 46 and 47 are very significantly lower than that of the corresponding E$_L$ region, but those of squares 54 and 55 are greater than the population of the corresponding left-handed region. The curvature of squares 54 and 55 is greater than that at squares 46 and 47, as is the torsion, and these structures are substantially closer to the A$_R$ region (we use capital A rather than α for the same reason that we used capital E above). Thus the gradual change in parity (handedness) preference from left- to right-handedness as we pass continuously along the positive κ axis from the E region, near the origin, to the A (helical) region is clearly manifest; i.e., near the origin, left-handedness predominates but, as we move away from the origin in the direction of increasing κ, right-handedness becomes increasingly preferred.

It should also be noted that island III is separated from island II by a channel of less-than-random occurrence (squares 31, 37, and 45) which falls in the extreme right-handed section of the E$_0$ region (opposite in handedness to squares 28, 34, and 43). This emphasizes again the strong preference for left-handedness[13–19] in the E region.

As was noted above, island III contains two local maxima, those at squares 46 and 62. The latter is the A$_R$-helical peak, which, as might be expected, overwhelms in magnitude any other feature of the distribution. The maximum at square 62 has associated with it squares 55, 61, and 63; these together are defined as the A$_R$ region. (A general procedure for dividing islands containing more than one maximum into separate peaks is outlined in Appendix II.)

Because of discontinuity no. 1 in the (κ, τ) plane, noted above, island I is actually contiguous structurally with the A$_R$ region of island III (see also Appendix I) in the sense that four-C$^\alpha$ structures represented by points which fall in square 7 are very similar to those represented by points which fall in square 63 (see paper 2 for details). It is thus clear that a continuum of bend structures exists, as was shown in paper 2. (We denote the bend structures that occupy island I, which are more planar than those of the A$_R$ region and include essentially flat bends near $\tau = 0$, as A$_0$ bends, in the same spirit as the E$_0$ notation above.) In fact, island I is contiguous also with the A$_L$ region (squares 9 and 57); however, the A$_L$ region is less-than-randomly occupied (except for the borderline square 51, which was discussed above).

The outcome of this analysis is a remarkable result. The region of greater-than-random occupation *forms a continuum* (broken by only one gap) *which includes four-C$^\alpha$ structures representative of the entire occupied region* (and therefore presumably the entire accessible conformational region) *except for the left-handed helical region.* This suggests that nucleation sites can be assembled out of four-C$^\alpha$ "pieces" representing almost the full range of

Table II
Characteristics of Peaks in Single-Site Distribution[a]

peak type	coordinates of maximum	contracted numbering of maximum	contracted numbering of squares in peak	population	$\bar{d}$	$\bar{s}$
A_0	(6, 8)	7	7–9	232	0.72	1.65
$E_L + E_0$	(11, 7)	42	(41–43, 51, 52)[b] (28–30, 34–36)[c]	1381 (552,[b] 829[c])	1.79	2.89
E_R	(11, 12)	46	46, 47, 54	223	0.72	1.47
A_R	(13, 15)	62	55, 61–63	1209	0.52	9.19

[a] All data pertain to Figure 2. [b] E_L region. [c] E_0 region.

possible structures. There are, of course, significant differences in population between different types of structure within the region of greater-than-random occurrence. This provides an important measure of discrimination between different types of nucleation sites in the sense that certain structural types, which are observed as maxima (e.g., E_L, E_R, and A_R), can be regarded as prototypical nucleation structures on the four-C^α length scale. Nevertheless, in view of the almost continuous distribution of frequently occurring structural types, it is clear that selectivity of nucleation is rather weak on the four-C^α length scale. It will therefore be of great interest to see what restrictions on pairing of four-C^α structures operate in the formation of structure on the five-C^α length scale. This problem will be taken up in the next section.

It should be noted that this result is independent of the size of the grid. An analysis of the distribution shown in Figure 2 using a grid twice as fine [(0.05 × 0.05) rad/Å] gives an essentially identical high-frequency region.

It is desirable to have a quantitative method for the representation of various properties of the peaks observed in bar graphs. For this purpose, we define the average value of a parameter D (which has the value D_i in square i) for a given peak as

$$\bar{D} = \sum_j D_j n_j / \sum_j n_j \qquad (1)$$

where the sum runs over all squares which are included in the peak and n_j is the population of the jth square.[20]

We shall use two parameters to characterize the peaks appearing in Figure 2. They are $\bar{d}$, the average distance of squares from the peak maximum, and $\bar{s}$, the average degree of nonrandomness of the peak. Each square in the 18 × 18 grid of Figure 2 can be characterized by its row and column position. Thus, square 3 has coordinates (5, 7) and square 61 has coordinates (13, 14). The distance d_{12} between two squares, say (m_1, n_1) and (m_2, n_2) is given by the usual Cartesian formula:

$$d_{12} = [(m_1 - m_2)^2 + (n_1 - n_2)^2]^{1/2} \qquad (2)$$

where d_{12} is measured in units of 0.1 rad/Å, the "lattice constant" of the graph. If one of these squares is the peak maximum, and the appropriate average is taken as in eq 1, $\bar{d}$ is obtained.[21] Clearly $\bar{d}$ is a measure of the width of the peak. Small values of $\bar{d}$ correspond to narrow peaks, and large values to broad peaks. It should be noted that s_i, the degree of nonrandomness of square i, is given by

$$s_i = n_i / 52.9 \qquad (3)$$

and that $s_i > 1$ corresponds to greater-than-random occurrence and $s_i < 1$ to less-than-random occurrence. Substitution of s_i of eq 3 for D_j in eq 1 yields $\bar{s}$.

The parameters calculated for the peaks observed in the three islands of Figure 2 are shown in Table II. The $E_L + E_0$ peak (island II) is clearly the broadest, as indicated by its large value of $\bar{d}$, 1.79. The A_0 and E_R peaks have essentially the same width, with $\bar{d} = 0.72$. The narrowest peak is the A_R peak, with $\bar{d} = 0.52$. (We stress again the fact that the A_R and A_0 peaks are structurally contiguous and that our separation of them is convenient, but somewhat arbitrary *on the four-C^α length scale*. In section III, we shall see that, on the five-C^α length scale, A_R and A_0 structures exhibit rather different behavior.) These values of $\bar{d}$ reflect accurately the fact that the α_R helix is the most sharply defined ordered backbone feature and the wide latitude which obtains in extended structures. The A_R peak has $\bar{s} = 9.19$, reflecting the fact that the A_R structure, by virtue of its occurrence either alone, as a bend, or in a combination with other A_R structures as α_R helices,[22–24] is overwhelmingly the most likely single, sharply defined structure to occur on the four-C^α length scale; it is followed in descending order of $\bar{s}$ by the $E_L + E_0$, A_0, and E_R peaks.

We can summarize the conclusions that we have drawn from the single-site (four-C^α structure) distribution as follows:

1. The structures which occur with greater-than-random frequency on this length scale form a virtual continuum, with only one break (in the channel formed by squares 31, 37, and 45 in the right-handed extended region), which includes structures representative of the entire occupied region of the (κ, τ) plane, except for the left-handed helical region.

2. We postulate that all of these high-frequency structures on the four-C^α scale are available to act as components of nucleation regions in the folding of the protein. It therefore seems that there is weak selection of nucleation structures on the four-C^α length scale and that stronger selection must take place on longer length scales in the protein.[8,25,26]

3. The extended region is seen to include four-C^α structures with either right- or left-handed twists, as well as many structures which are almost planar. These are designated as E_R, E_L, and E_0 structures, respectively. The preference for left-handedness in the extended region is clearly demonstrated.[13–19]

4. Parameters were developed to describe quantitatively the localization and population of the various peaks observed in the single-site (four-C^α) distribution. It was shown that the A_R peak is very narrow and very high, that the $E_0 + E_L$ peak is quite broad and rather high, and that there are two lower peaks which are narrow—the A_0 (flat bend) and E_R peaks.

We now extend the methods that we have developed to the study of the nearest-neighbor pair distribution of values of (κ, τ), which reflects the structure of the backbone on the *five-C^α* length scale.

III. Nearest-Neighbor Distribution

The protein sample given in Table I leads to 3729 nearest-neighbor pairs[27] of values of $(\kappa_i, \tau_i; \kappa_{i+1}, \tau_{i+1})$, the distribution of which can be represented on a graph in which are given the populations of (0.1 × 0.1 × 0.1 × 0.1) rad/Å hypercubes in the $(\kappa_i, \tau_i; \kappa_{i+1}, \tau_{i+1})$ hyperplane. These four values, plus the height, constitute essentially a five-dimensional bar graph. It should be noted that it is also possible to represent the data in a three-dimensional contracted graph by assigning to each hypercube two coordinates corresponding to the arbitrary successive numbering that we have assigned to the occupied squares in the (κ, τ) plane (Figure 2). We shall find this to be a convenient shorthand notation for identifying hypercubes. The contraction, however, eliminates the structural information contained in the five-dimensional representation, so that we cannot use it for structural studies; these must be carried out in the full five-dimensional space.

We shall find it convenient to have a notation which enables us to distinguish succinctly between the (κ, τ) and (κ_i, τ_i; κ_{i+1}, τ_{i+1}) planes and between these and the contracted representation of the latter. We will therefore refer to these henceforth as $\mathbf{P}^{(2)}$, $\mathbf{P}^{(4)}$, and $\mathbf{C}^{(2)}$, respectively. The single-site (three-dimensional) and nearest-neighbor (five-dimensional) graphs will be denoted by $\mathbf{G}^{(3)}$ and $\mathbf{G}^{(5)}$, respectively.

The fact that we are forced to operate in a five-dimensional space deprives us of the advantages afforded by geometrical intuition. We must therefore fall back on analytical methods which, in the three-dimensional case, reduce to intuitively evident procedures.

The first step in our analysis is to locate the maxima in $\mathbf{G}^{(5)}$. This is done by searching for hypercubes in $\mathbf{P}^{(4)}$ whose population is greater than that of any of their nearest neighbors. (We note in passing that the number of nearest neighbors in n dimensions is given by $3^n - 1$, so that, whereas a given cube in $\mathbf{P}^{(2)}$ has 8 nearest neighbors, a hypercube in $\mathbf{P}^{(4)}$ has 80.) In order to decide which of these peaks are of sufficient height to be significant, we observe that the 3729 pairs are distributed among only 1052 of the 5041 ($=71^2$) theoretically accessible hypercubes, where 71 is the number of occupied squares in the single-site distribution of Figure 2. If the pairs were distributed randomly in this occupied subspace, there would be a uniform population per hypercube of 3729/1052, or 3.54 (≈ 4). We shall therefore accept as significant structural features those peaks with height $h \geq 5$ and assign hypercubes to them by using an exactly analogous procedure (Appendix II) to that outlined in section II, with the proviso that the population n_i of such hypercubes be greater than 4 (to distinguish it from random occupancy).[12] This procedure for constructing peaks may err on the side of conservatism in neglecting hypercubes with $n_i < 4$; nevertheless, it seems to be a reasonably satisfactory method for picking peaks out of the background.

Another factor that must be considered is that of correlation. It should be noted that, in constructing the nearest-neighbor graph $\mathbf{G}^{(5)}$, we use a sample which also has the single-site distribution given by $\mathbf{G}^{(3)}$ (Figure 2). Thus, the degree of randomness of pairing, or degree of correlation, is given by

$$c_{ij} = s_{ij}/s_i s_j \qquad (4)$$

where $s_k = n_k/3754$ is the fraction of the single-site distribution occurring in square k and $s_{kl} = n_{kl}/3729$ is the fraction of nearest-neighbor pairs occurring in hypercube kl (where we use the contracted numbering of Figure 2). $c_{ij} > 1$ indicates that the nearest-neighbor pair ij occurs in $\mathbf{G}^{(5)}$ with greater frequency than would be expected on the basis of the frequencies of i and j in $\mathbf{G}^{(3)}$; $c_{ij} < 1$ indicates the reverse.

As in the case of $\mathbf{G}^{(3)}$, we shall be interested in the distance d of a given hypercube in $\mathbf{P}^{(4)}$ from the maximum with which it is associated. A new parameter of interest in the analysis of $\mathbf{G}^{(5)}$ is the distance Δ_{ij} *in* $\boldsymbol{P}^{(2)}$ between the two squares which make up the nearest-neighbor pair. The smaller this distance is, the more nearly the hypercube in question represents a (regular) repeating structure. In analogy with eq 1, these parameters can all be used in the form of population-weighted averages over the hypercubes in a given peak:

$$\bar{a} = \sum_{ij} n_{ij} a_{ij} \Big/ \sum_{ij} n_{ij} \qquad (5)$$

In using these parameters, account must be taken of the mathematical features of the (κ, τ) representation summarized in the previous section. This is particularly true because, as in the discussion of $\mathbf{G}^{(3)}$ in section II, and for the reason noted there, we will deal with the untransformed nearest-neighbor distribution characterized by mixed values of $p_i p_{i+1}$ ($=\pm 1$). [In Appendix III we demonstrate that the transformed ($p_i p_{i+1} = -1$ only) nearest-neighbor distribution gives results which are completely consistent with the analysis in the main text.]

Of the 32 maxima found in $\mathbf{G}^{(5)}$, there are 20 with $h \geq 5$. These fall on 11 islands composed of hypercubes with population greater than 4; the peaks are shown, organized by island, in Table III. There is one additional small island (no. XII of Table III) with a double peak, i.e., two adjacent hypercubes with the same population. As in the discussion of $\mathbf{G}^{(3)}$, we remark that the total number of hypercubes contained in these islands is 164, or 15.6% of the total number of occupied hypercubes, whereas the total population of these islands is 2219 nearest-neighbor pairs, or 59.5% of the total number of pairs. This indicates clearly that the usefulness of the island approach for the determination of significant structural features is retained in $\mathbf{G}^{(5)}$.

Table III
Peaks in $\mathbf{G}^{(5)}$ Organized by Islands

island	coordinates of peak[a]	contracted coordinates	height h
I	(6, 8, 6, 8)	(7, 7)	8
II	(6, 8, 13, 15)	(7, 62)	19
	(6, 9, 11, 12)	(8, 46)	7
III	(7, 8, 9, 9)	(13, 29)	6
IV	(7, 9, 11, 7)	(14, 42)	8
	(8, 12, 11, 7)	(25, 42)	9
	(9, 9, 11, 6)	(29, 41)	20
	(10, 8, 9, 9)	(34, 29)	31
	(11, 6, 11, 7)	(41, 42)	16
V	(10, 9, 7, 8)	(35, 13)	7
VI	(10, 9, 13, 15)	(35, 62)	22
	(11, 7, 13, 15)	(42, 62)	25
	(11, 12, 13, 16)	(46, 63)	8
	(13, 15, 13, 16)	(62, 63)	261
VII	(12, 13, 11, 12)	(54, 46)	6
VIII	(11, 7, 11, 12)	(42, 46)	8
IX	(11, 12, 11, 7)	(46, 42)	9
	(13, 16, 12, 6)	(63, 52)	9
X	(13, 15, 6, 8)	(62, 7)	28
XI	(13, 15, 8, 12)	(62, 25)	7
XII	(12, 13, 9, 8)	(54, 28)	8[b]
	(12, 14, 10, 9)	(55, 35)	8[b]

[a] Row and column numbering from Figure 2. [b] Double peak (see text).

The peaks occurring in $\mathbf{G}^{(5)}$ can be classified according to the types of structure they represent, using as a basis the classification of four-C$^\alpha$ units established in section II. The results are shown in Table IV.

We begin our discussion of Table IV by considering the EE peaks, which represent five-C$^\alpha$ structures which fall entirely in the extended region. It can be seen that nearly all possible combinations of the three extended types that we defined previously ($\mathrm{E_R}$, $\mathrm{E_L}$, and $\mathrm{E_0}$) occur. It is instructive to examine the relative importance of the dif-

Table IV
Peaks in $G^{(5)}$ Organized by Type

	peak type	coordinates of maximum	contracted coordinates of maximum	N_c[a]	h[b]	$\bar{c}$	$\bar{d}$	$\bar{\Delta}$	population of peak
EE	EE	(8, 12, 11, 7)	(25, 42)	8	9	3.55	1.31	4.77	51
	E_RE_L	(11, 12, 11, 7)	(46, 42)	6	9	2.43	1.22	6.22	41
	E_LE_R	(11, 7, 11, 12)	(42, 46)	2	8	2.12	0.54	5.41	13
	E_LE_L	(11, 6, 11, 7)	(41, 42)	16	16	2.03	1.43	1.57	128
	E_RE_R	(12, 13, 11, 12)	(54, 46)	1	6	3.26	0.00	1.41	6
	E_0E_L	(9, 9, 11, 6)	(29, 41)	17	20	2.39	1.29	2.93	163
	E_0E_0 + E_LE_0[c]	(10, 8, 9, 9)	(34, 29)	37	31	2.31	1.60	1.65	444
						(2.40, 2.04)[c]	(1.46, –)	(1.09, 3.24)	(328, 116)
	E_RE_0	(12, 13, 9, 8)	(54, 28)	2	8	4.40	0.77	4.80	13
EA_R	E_0A_R	(10, 9, 13, 15)	(35, 62)	8	22	0.98	0.81	6.34	63.5
	E_LA_R	(11, 7, 13, 15)	(42, 62)	9	25	0.80	1.00	8.83	81.5
	E_RA_R	(11, 12, 13, 16)	(46, 63)	5	8	0.84	1.03	3.66	32
A_RE	A_RE_0	(12, 14, 10, 9)	(55, 35)	1	8	1.45	0.00	5.39	8
	A_RE_L	(13, 16, 12, 6)	(63, 52)	4	9	0.69	0.80	9.23	28
	A_RE_R	(13, 15, 8, 12)	(62, 25)[d]	1	7	0.95	0.00	5.83	7
	E_0A_0	(10, 9, 7, 8)	(35, 13)[d]	2	7	2.51	0.59	7.97	12
A_0E	A_0E_R	(6, 9, 11, 12)	(8, 46)	5	7	4.15	1.12	5.47	29
	A_0E_L	(7, 9, 11, 7)	(14, 42)[d]	3	8	3.01	0.93	8.08	20
	A_0E_0	(7, 8, 9, 9)	(13, 29)[d]	6	6	3.04	1.34	7.59	31
	A_RA_R + A_RE_R[c]	(13, 15, 13, 16)	(62, 63)	19	261	2.68	1.04	0.86	912
						(2.78, 0.93)[c]	(0.9, –)	(0.74, 3.11)	(866, 46)
	A_RA_0	(13, 15, 6, 8)	(62, 7)	5	28	1.43	0.62	1.10	67
	A_0A_R	(6, 8, 13, 15)	(7, 62)	5	19	1.56	0.80	1.59	54
	A_0A_0	(6, 8, 6, 8)	(7, 7)	2	8	2.79	0.47	0.53	15

[a] Number of hypercubes in peak. [b] Height of maximum (from Table III). [c] See text for discussion of these peaks, which contain two types of structure. [d] Square 25 is structurally adjacent to the E_R region because of the second discontinuity mentioned in section II of the text. Squares 13 and 14 do not fall within the A_0 island of Figure 2 but are structurally adjacent to it.

ferent peak types. The first peak listed, with maximum at (25, 42) (contracted coordinates), contains a number of dissimilar types of extended structure, i.e., structures with different handedness (not shown explicitly in the table), and arises because of the mixing of $p_ip_{i+1} = +1$ and $p_ip_{i+1} = -1$ structures in this limited region of $G^{(5)}$. (For this reason, this peak is denoted only as EE in Table IV.) It will be seen from the analysis of the transformed version of $G^{(5)}$ (Appendix III) that the proper assignment of the contents of this peak does not affect any of the conclusions of this section. We also observe that the peak with maximum at (34, 29) contains two types of structure, E_0E_0 and E_LE_0, with respective populations of 328 and 116. The fact that these are contained in a single peak indicates that the two structural types form a continuum, with all intermediate structural types between the two extremes (E_0E_0 and E_LE_0) occurring in substantial numbers. This undoubtedly corresponds to the fact (noted in section II) that a similar continuum connecting E_L and E_0 types occurs in $G^{(3)}$. (The occurrence of two structural types in *separate* peaks indicates that structures intermediate in character between the two occur with *lesser* frequency than the two extreme structures.)

The populations of the structurally defined EE peaks (given in the last column of Table IV) are ordered as follows: (E_0E_0) > E_0E_L > E_LE_L > (E_LE_0) > E_RE_L > E_LE_R = E_RE_0 > E_RE_R. Two types are placed in parentheses to indicate that they occur in the same peak and therefore are divided by inspection into the two classes. It is seen from the above inequality that structures which are made up of various combinations of E_0 and E_L four-C$^\alpha$ units predominate in the extended region on the five-C$^\alpha$ scale. There is a strong preference for nearly flat E_0E_0 structures. E_0E_L structures are markedly preferred to E_LE_0 types, which occur with nearly the same frequency as E_LE_L structures. Structures containing E_R units are considerably less common. Within this group, the most frequently occurring structure is the E_RE_L type, which is heavily favored over the E_LE_R type. This latter occurs with the same frequency as the E_RE_0 structure, and the least common feature is the E_RE_R five-C$^\alpha$ unit.

It is of interest to note that, in mixed EE structures, E_L four-C$^\alpha$ units show a strong preference for the second position (E_XE_L), whereas E_R structures prefer the first position (E_RE_X), where X is R, L, or 0.

All of the peaks in the EE region have positive average correlation between the component four-C$^\alpha$ units ($\bar{c} > 1$), i.e., they pair with greater-than-random frequency, confirming the tendency of E units to associate to form extended strands. Although E_R structures occur least frequently, relative to their population, they are the most likely to associate with other E_X structures; E_0 and E_L (each associated with other E_X) structures occur even less frequently relative to their populations. The width of the peaks, as measured by $\bar{d}$, varies from 0 for the E_RE_R peak at (54, 46), which includes a single hypercube, to 1.60 for the E_0E_0 + E_LE_0 peak at (34, 29). As indicated in Table IV, it is possible to calculate separate values of the various parameters for the two components of this peak (given in parentheses). The value of $\bar{d}$ for the E_LE_0 component is not meaningful, because the peak contains two kinds of structures, and the maximum is not contained in the E_LE_0 region of the peak. The value of $\bar{d}$ for the E_0E_0 component, however, is meaningful; it is 1.46, making this the broadest peak in $G^{(5)}$. This is followed closely by the E_LE_L peak at (41, 42) with $\bar{d}$ = 1.43. These peaks therefore contain a relatively broad range of structural types. The peaks then narrow (i.e., decrease in $\bar{d}$) in the order E_0E_L, E_RE_L, E_RE_0,

E_LE_R, E_RE_R. The large difference in $\bar{d}$ between the E_RE_L and E_LE_R peaks corresponds to their difference in population.

The most frequently occurring structures, those constituting the E_0E_0 component of the peak at (34, 29), show a strong tendency to be repeating, with $\bar{\Delta} = 1.09$. This is also true of the E_LE_L peak, with $\bar{\Delta} = 1.57$, and of the very small E_RE_R peak, with $\bar{\Delta} = 1.41$. The E_LE_0 and E_0E_L peaks, which represent transitions between twisted and flat structures, have intermediate values of $\bar{\Delta}$ at 3.24 and 2.93, respectively. The small E_RE_0 peak is somewhat less repeating, with $\bar{\Delta} = 4.80$. The E_RE_L and E_LE_R peaks, which mix twisted four-C^α structures of opposite handedness, have values of $\bar{\Delta}$ of 6.22 and 5.41, respectively. We see then that extended structures which are almost repeating predominate; this is in keeping with the prevalent intuitive picture of an extended strand. There are, however, substantial numbers of extended structures which are not repeating structures.

The next two groups of peaks in Table IV represent five-C^α structures in which the various types of extended four-C^α units combine with four-C^α segments which fall in the A_R region. Peaks corresponding to all possible E_XA_R and A_RE_X structures are observed. The populations of the three EA_R peaks are ordered as $E_LA_R > E_0A_R > E_RA_R$, with a marked preponderance of E_LA_R over the other two possibilities. The E_0A_R and E_LA_R peaks are rather narrow ($\bar{d} = 0.81, 1.00$). The E_RA_R peak is less populated but of equal width ($\bar{d} = 1.03$). The values of $\bar{\Delta}$ are reasonable for structures which are clearly not repeating. The most nearly repeating structure is contained in the E_RA_R peak, with $\bar{\Delta} = 3.66$. The most striking feature of these peaks is the fact that, for almost all of them, $\bar{c} \lesssim 1$, indicating that (on the five-C^α length scale) *E and A_R structures tend to avoid one another*. [The sole exception is the small A_RE_0 peak at (55, 35), which is actually distorted from the classical A_R region toward an E_RE_0 type of structure.] This is reasonable on steric grounds. One expects that the primary circumstance under which E and A (bend) structures pair should be at turns involved in β-sheet structures. Clearly, A_R bends, which are quite nonplanar, are not favored under conditions where it is necessary to ensure the correct registration of the extended strands on either side of the bend.

The same effect is observed in the group of A_RE peaks. This includes the three peaks listed under the A_RE heading, as well as the A_RE_R component of the enormous A_RA_R peak, which is separated by inspection in the same manner as the $E_0E_0 + E_LE_0$ peak above. The values of $\bar{c}$ of these are less than 1, with the exception of the A_RE_0 peak, as noted above. It should be noted that the number of A_RE structures is considerably less than the number of EA_R structures and that the peaks are narrow.

In contrast to the above observations, the values of $\bar{c}$ of the A_0E_X and E_0A_0 peaks are all quite large. It is clear that flat bends prefer to associate with extended structures, probably because of the role that these bends play in the construction of β structure. The A_0E peaks are all broader ($0.93 \leq \bar{d} \leq 1.34$) than the narrow ($\bar{d} = 0.59$) E_0A_0 peak. This correlates with the marked asymmetry in population, with A_0E structures being much more common. It is also of interest that E_RA_0 and E_LA_0 structures occur with very low frequency—so low, in fact, that corresponding peaks do not appear in $\mathbf{G}^{(5)}$.

The remaining peaks in $\mathbf{G}^{(5)}$ involve combinations of the two bend types. The A_RA_R peak at (62, 63) involves a minor A_RE_R component, discussed above. The enormous helical component has positive correlation ($\bar{c} = 2.78$), very sharp definition ($\bar{d} = 0.9$) considering its population (866), and is, of course, a repeating structure ($\bar{\Delta} = 0.74$).

The two peaks involving combinations of the A_0 and A_R four-C^α structures show positive, but rather low, correlation. Clearly, these structures are substantially less favored than A_RA_R structures, despite the fact that they are related structurally rather closely. A measure of this relation is given by the values of $\bar{\Delta}$ [1.10 for the A_RA_0 peak at (62, 7) and 1.59 for the A_0A_R peak at (7, 62)], which indicate that these structures are substantially less repeating than A_RA_R ($\bar{\Delta} = 0.74$). Both peaks are fairly narrow, with $\bar{d} = 0.62$ (A_RA_0) and 0.80 (A_0A_R).

The last peak in $\mathbf{G}^{(5)}$ is the A_0A_0 peak at (7, 7), which shows substantial positive correlation ($\bar{c} = 2.79$). The peak is quite narrow ($\bar{d} = 0.47$), and the $\bar{\Delta}$ value of 0.53 indicates that the points in this peak represent quite strongly repeating structures.

The foregoing results indicate that there are substantial differences in the behavior of A_R and A_0 four-C^α structures on the five-C^α length scale, despite the fact that they are rather closely related structurally. In fact, they seem to exhibit complementarity in their interactions with neighboring structures; in a given situation, one type of bend is preferred, and the other tends to be avoided. In combination with extended structures, A_R bends exhibit negative correlation, and A_0 structures positive correlation. In combination with other A_R structures, A_R shows highly positive, and A_0 substantially less positive, correlation. We see, then, that the selectivity which seemed to be absent, or very weak, in the choice of possible nucleating structures on the four-C^α length scale begins to become more marked on the five-C^α scale. Consideration of the four-C^α structures contained in $\mathbf{G}^{(3)}$ revealed that any conformation accessible to a four-C^α unit is similar to some structure which occurs with greater-than-average frequency and is therefore convertible with minimal deformation into a potential nucleating structure on the four-C^α scale. Consideration of $\mathbf{G}^{(5)}$, however, reveals that there are restrictions on the manner in which these four-C^α nucleation structures can coalesce. Of greatest significance is the fact that two major four-C^α nucleation structures, the E and A_R structures, *show a pronounced tendency to avoid one another*. The structure which shows a positive tendency to associate with both, the A_0 structure, occurs less frequently.

From the foregoing considerations, a hypothesis about protein folding can be formulated. As renaturing conditions are created, nuclei of relatively long-lived structure are formed. These begin as very local, single-residue structures and lengthen as conditions become more renaturing. These structures are likely to be those which are most strongly stabilized by local interactions and therefore also appear with highest frequency in the native structure.

Our analysis suggests that the nucleating structures which form *up to the five-C^α length scale* fall into *either* of two classes: (a) essentially repeating structures, which can be either extended strands or right-handed helices, or (b) nonrepeating structures. (As pointed out above, extended strands are not *strictly* repeating structures, being composed of E_L, E_R, and E_0 four-C^α units.) This is a more general result than is suggested by some previous work,[8,9] in which the nuclei are assumed to fall into one or the other class. Nonrepeating structures are likely to involve A_0 (flat) bends in combination with either E_X or A_R structures. Nucleating structures which are combinations of A_R and E_X four-C^α structures are not likely to occur.

We have said nothing about the order in which various nuclei form in the chain, and indeed our analysis does not

Table V
Peaks in Transformed $G^{(5)}$

peak type		coordinates	contracted coordinates	N_c	h	$\bar{c}$	$\bar{d}$	$\bar{\Delta}$	population
EE	E_RE_L	(11, 12, 11, 7)	(46, 42)	6	9	3.05	1.18	6.24	39.5
	E_LE_R	(12, 4, 11, 12)	(50, 46)	1	5	4.23	0.0	8.06	5
	E_LE_R	(11, 7, 11, 12)	(42, 46)	2	8	2.13	0.54	5.42	13
	E_LE_L	(11, 6, 11, 7)	(41, 42)	19	16	2.03	1.49	1.77	147
	E_RE_R	(12, 13, 11, 12)	(54, 46)	1	6	3.25	0.0	1.41	6
	E_0E_L	(9, 9, 11, 6)	(29, 41)	25	20	2.70	1.65	3.34	210
	$E_0E_0 + E_LE_0$	(10, 8, 9, 9)	(34, 29)	37	31	2.32	1.60	1.66	445
	E_RE_0	(12, 13, 9, 8)	(54, 28)	2	8	4.40	0.77	4.80	13
EA	$E_0A_R + E_0A_0$	(10, 9, 13, 15)	(35, 62)	13	22	1.39	1.04	6.61	94
	E_LA_R	(11, 7, 13, 15)	(42, 62)	12	26	0.96	1.16	9.0	111
	$E_RA_R + E_RA_0$	(11, 12, 13, 16)	(46, 63)	6	10	1.03	0.99	3.88	41
A_RE	A_RE_0	(12, 14, 10, 9)	(55, 35)	1	8	1.45	0.0	5.39	8
	A_RE_R	(13, 15, 8, 12)	(62, 25)	1	7	1.32	0.0	5.83	7
	$A_RE_L + A_RE_0 + A_0E_L + A_0E_0$	(13, 16, 11, 7)	(63, 42)	12	13	1.44	1.69	8.57	80.5
	$A_RA_R + A_RE_R + A_RA_0 + A_0A_0 + A_0A_R$	(13, 15, 13, 16)	(62, 63)	29	286	2.36	1.15	1.0	1111

provide any information on this point. It should be noted, however, that none of the structures which we have proposed as five-C$^\alpha$ nucleating structures can be ruled out as the initial nucleating structure. At first glance it may seem unlikely that the formation of an extended strand, for example, should be a necessary first step in the refolding process. It is conceivable, however, that the *separation* of two parts of the molecule, which can be accomplished through the formation of a linear structure, is a necessary preliminary to correct folding in some proteins. In other cases, the formation of one or several bends may be the initial step in refolding. We thus take a somewhat more general view of nucleation than did Matheson and Scheraga,[8] who postulate hairpin bend formation as the initial step in nucleation. This difference may be more semantic than real, however, since those authors based their work on hydrophobic interactions and therefore *defined* the primary (initial) nucleating structure as one which is stable and promotes a more compact chain structure, with strong hydrophobic bonding, irrespective of any repeating structure which may be present. Here, we take the view that a decrease in chain dimensions is not necessarily the *first* step in correct folding and therefore permit the consideration of repeating structures as initial nuclei. The (sometimes) larger structures of ref 8 form after those on the five-C$^\alpha$ length scale.

Similarly, our views are more general than those of Tanaka and Scheraga,[25] who postulate that the initial folding nuclei are those (including α helices[9]) which appear near the diagonal in the protein contact map. It should be emphasized again that none of these viewpoints are excluded by our analysis. We maintain that the *kinetically visible* nucleating step, for example, leads to a properly folded native conformation *only because of* interactions brought about by other locally stabilized structures present in the chain. We therefore regard these structures as nuclei essential for proper folding. In this sense, the differential geometric analysis provides a unifying framework for the various proposed nucleation mechanisms. It also illustrates that selectivity of nucleation types increases as nucleation proceeds to longer length scales. Further, it is able to identify precisely the types of nonregular structure[8,25,26] that are likely to take part in nucleation, and which nonregular structures are apparently not sufficiently stabilized by purely internal interactions, and are therefore stabilized by interactions on a larger length scale (e.g., those discussed in ref 8) than that which we have explored thus far.

An interesting question raised by these results is whether there is an upper limit to the length scale on which characteristic structures can be observed which are common to a wide assortment of proteins. Intuitively one expects that such a limit does exist. Inspection of structure on a scale above this limit should reveal a broad distribution of structural types which reflect the individual folding patterns of the various proteins (or classes of proteins) in the sample. We have no quantitative answer to this question at present.

Another question of interest is the influence of residue type on local chain structure as revealed by the differential geometric representation. The results presented above are averaged over the amino acid sequence; one expects that there will be variations as a function of amino acid composition. This problem will be addressed in a forthcoming publication.

IV. Summary

We have used a graphical method based on the differential-geometric approach to investigate the backbone structure of proteins on the four- and five-C$^\alpha$ length scales. The result of this investigation led to a number of conclusions about protein structure and protein folding.

1. On the four-C$^\alpha$ length scale, certain distinct structural types are observed. The extended region shows two peaks, one corresponding to right-handed twisted four-C$^\alpha$ structures (E_R) and one containing left-handed twisted (E_L) structures and nearly flat (E_0) structures. There is a structural continuum embracing the latter two, which is separate from the smaller E_R peak. Another structural continuum observed on the four-C$^\alpha$ scale involves the very common α-helical (A_R) and less common nearly flat (A_0) bend types.

2. An analysis of the types of structure which occur with high frequency reveals that such structures are distributed *throughout the occupied region* on the four-C$^\alpha$ length scale. Therefore, any structure accessible to a four-C$^\alpha$ unit is close to some such structure, probably within fluctuation range. It seems reasonable that such high-frequency structures are energetically favored and are therefore potential nucleation structures in the denatured protein. Thus, potential nucleation structures are distributed throughout the accessible region on the four-C$^\alpha$ length scale, and there is rather weak selectivity in nucleation on this length scale.

3. Extension of this analysis to five-C$^\alpha$ structures reveals distinct peaks corresponding to various combinations of the four-C$^\alpha$ structures noted in (1) above. The extended region is particularly rich in peak structure, with peaks corresponding to most of the possible combinations of E_L,

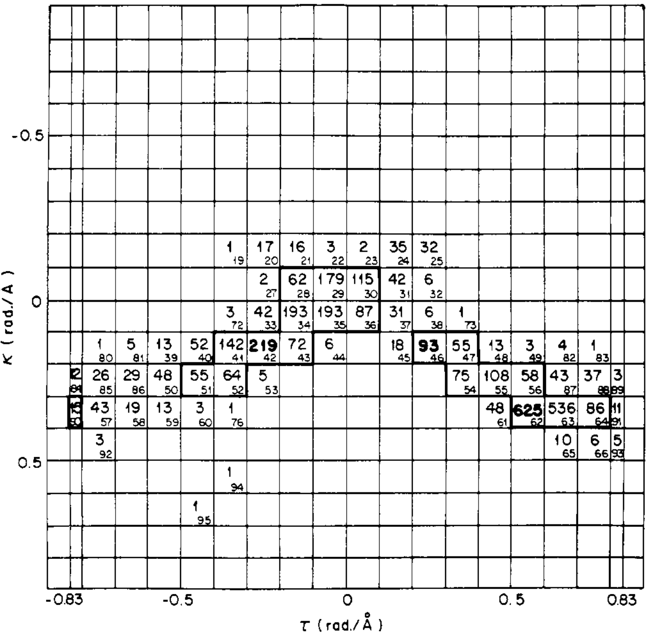

Figure 3. Transformed single-site distribution of values of (κ, τ) for the same protein sample. Notations as in Figure 2. It should be noted that the lines $\tau = \pm 0.83$ are equivalent because of a cyclic boundary condition. The numbering of the squares (small numbers) is the same as in Figure 2; squares with numbers greater than 76 are not occupied in the untransformed distribution of Figure 2.

E_R, and E_0 structures. There are wide variations in the frequency with which these structures occur. Other peaks are observed corresponding to combinations of extended structures with A_0 and A_R bend types, to A_R-helical segments, to combinations of A_0 and A_R bends and to A_0A_0 structures. Analysis of the correlation between neighboring structures reveals that there is a tendency for certain structures to avoid association and for others to associate preferentially. This leads to complementarity between similar structural types. For example, the two bend types are complementary in their association with extended structures; A_0 bends tend to associate with extended four-C$^\alpha$ structures and A_R bends to avoid them.

4. We suggest that those five-C$^\alpha$ structures which exhibit positive correlation between their component four-C$^\alpha$ structures are potential five-C$^\alpha$ components of nucleating structures. These are the E_XE_Y, E_0A_0, A_0E, A_RA_R, and A_0A_0 structures. A_RA_0 and A_0A_R show positive, but considerably lower, correlation. Clearly there is a higher level of selectivity in nucleation on the five-C$^\alpha$ length scale, brought about by the preferential association of certain structures. There is also a considerable difference in the roles played by the various potential nucleation structures. The EE and A_RA_R structures seem to provide nuclei for the formation of repeating extended or helical structures, respectively. The role of the A_0 structure is important despite its numerical inferiority, since it seems to provide the principal building block (in combination with either E or A_R units) for nucleation of nonregular backbone structures, which are crucial for the formation of the final globular conformation of the protein molecule.

Acknowledgment. We thank Dr. George Némethy and Professor Michael Cowen for helpful discussions.

Appendix I. Analysis of Transformed Single-Site Distribution

In Figure 3, we show the transformed version of Figure 2; i.e., all points of Figure 2 with $p_ip_{i+1} = +1$ have been transferred (in Figure 3) to squares with $p_ip_{i+1} = -1$ by the methods outlined in paper 2.[5] An analysis exactly analogous to that performed for Figure 2 in the main text gives two islands of greater-than-random occupancy, which are indicated in the figure by heavy outlines. It can be seen (by comparing Figures 2 and 3) that the island including the E_L and E_0 structures remains unaltered. The island containing the E_R and A_R peaks is now expanded, due to the transposition of the A_0 region to positions adjacent to the A_R and A_L regions. It should be noted that square 64, which was not part of this island in Figure 2, is now included (since it now contains points which fell in the A_0 region in Figure 2) and the population of square 63 is considerably increased. It should also be noted that the partial square 90 is also a region of marginally high frequency, corresponding to square 9 of Figure 2. It can thus be seen that the same structures which appear in the high-frequency islands of Figure 2 appear in those of Figure 3. Figure 3 also brings out the close structural relationship between the A_R, A_0, and A_L regions. (It should be noted that in Figure 3 there is a periodic boundary condition operating between the lines $\tau = \pm 0.83$, so that, for example, "squares" 90 and 91 are adjacent.)

Appendix II. Procedure for Separating Islands into Peaks

The following procedure is valid in any dimension and reduces in the three-dimensional case to an intuitively clear procedure.

(1) For each maximum on the island, all squares are identified which are accessible from the maximum by proceeding downward or horizontally (in population). In general, there will be squares which are accessible from more than one peak. Therefore, (2) each square which is accessible from more than one peak is assigned to the peak to which it is closest; except that (3) squares which are nearest neighbors of two peaks simultaneously (in a sense to be defined) are divided equally between the two peaks.

The sense in which we define nearest neighbors is that two squares, for example, (m_1, n_1) and (m_2, n_2) in a three-dimensional graph, are nearest neighbors if

$$\max\,(|m_1 - m_2|, |n_1 - n_2|) = 1$$

As pointed out in section III of the main text, each square in $\mathbf{P}^{(2)}$ has 8 nearest neighbors in this sense, whereas each hypercube in $\mathbf{P}^{(4)}$ has 80 nearest neighbors.

Appendix III. Analysis of Transformed Version of $\mathbf{G}^{(5)}$

Table V shows the peaks in the transformed version of $\mathbf{G}^{(5)}$ organized by type, in analogy with Table IV. It can be seen that the essential change in the extended region brought about by the transformation is the redistribution of the contents of the mixed peak [at (25, 42) of Table IV], principally to the E_LE_L peak at (41, 42) and the E_0E_L peak at (29, 41). This increases the values of $\bar{d}$ and $\bar{\Delta}$ of these peaks somewhat. The conclusions drawn in the main text remain unaffected.

The principal effect of the transformation is to place the A_0 region next to the A_R and A_L regions of the (κ, τ) plane. This results in the elimination from Table V of separate peaks containing the A_0 structures. The peaks centered at the A_R region are broader and contain A_0 structures as well as A_R structures. The values of $\bar{c}$ of these peaks are intermediate between those of the corresponding separate peaks in Table IV, reflecting the complementarity between A_R and A_0 structures noted in the main text (section III).

It is clear that the use of the untransformed version of $\mathbf{G}^{(5)}$ enables us to separate the roles of A_R and A_0 structures conveniently, with no loss of information.

References and Notes

(1) This work was supported by research grants from the National Science Foundation (Grant No. PCM79-20279) and from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (Grant No. GM-14312).

(2) (a) NIH Postdoctoral Fellow, 1977–1978; Todd Postdoctoral Fellow, 1978–1979. (b) Author to whom requests for reprints should be addressed at Cornell University.

(3) Venkatachalam, C. M. *Biopolymers* **1968**, *6*, 1425.

(4) Lewis, P. N.; Momany, F. A.; Scheraga, H. A. *Biochim. Biophys. Acta* **1973**, *303*, 211.

(5) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1440.

(6) Nêmethy, G.; Scheraga, H. A. *Biochem. Biophys. Res. Commun.* **1980**, *95*, 320.

(7) Rose, G. D. *Nature (London)* **1978**, *272*, 586.

(8) Matheson, R. R., Jr.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 819.

(9) Finkelstein, A. V.; Ptitsyn, O. B. *J. Mol. Biol.* **1976**, *103*, 15.

(10) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 1168.

(11) By definition,[5,10] $p_ip_{i+1} = +1$ for A_0 bends.

(12) In order to assess the significance of the observed difference between the three- and five-dimensional distributions (of parts II and III, respectively) and the uniform distributions to which they are compared in the text, χ^2 goodness-of-fit tests were carried out. In both cases it was determined that the distributions differ from the corresponding uniform distribution with greater than 0.999 significance.

(13) It should be noted that the discussion of the parity (i.e., handedness) of the extended region in paper 1 is in error, due to the error in the assignment of parity regions in that paper; a corrected presentation is given in paper 2. In fact, it is known[14,15] that, while β sheets have predominantly right-handed twists, the extended strands from which they are constituted are mainly left-handed in their twist. Our observations in this paper are in agreement with this earlier conclusion.[14,15]

(14) Chothia, C. *J. Mol. Biol.* **1973**, *75*, 295.

(15) There has been some ambiguity in the literature concerning the definition of the handedness of individual strands in β sheets. We define handedness in a manner consistent with the nomenclature for the conformations of polypeptide chains;[16] i.e., the sign of the dihedral angle around a *virtual* bond is defined in the same way as that around a *chemical* bond in Figure 1 of ref 16. With this definition, all of the *individual strands* in the β sheets commonly observed in proteins have a predominantly left-handed twist. Our definition of parity is in agreement with this definition of handedness. Also, this definition of handedness is consistent with that derived earlier[17] from an analysis of curves of constant n (where n is the number of residues per turn of helical structures) on a (ϕ, ψ) diagram: the dihedral angles of the individual strands in the β sheets observed in proteins fall predominantly in the region of the (ϕ, ψ) map that corresponds to left-handedness.[14,18] Furthermore, if one looks at *successive* backbone oxygen atoms of these individual strands, the line connecting them forms a *left*-handed helix around the backbone. If one considers *every second* backbone oxygen atom, however, then these form a *right*-handed helix around the backbone. The direction of hydrogen bonding between two adjacent strands in a β sheet is determined primarily by the directions of every second C=O bond of each strand, and hence the curved sheet passing through the successive hydrogen bonds between two strands in the commonly observed β structures in proteins has a *right-handed* twist. In other words, left-handed β strands combine to form a β sheet with a right-handed twist, when viewed along the strands. This relationship between the twist of the strands and of the sheet has been demonstrated in Figure 4 of ref 18, which used the same convention as that adopted here. An analogous relationship is seen in collagen and collagen-like triple-stranded polypeptides: the individual strands form left-handed helices, but the triple-stranded structure has a right-handed twist around its major axis.[19] In several other studies,[14] the orientation of *every second* residue, rather than of every residue, has been used to define the twist of the *individual strands*, leading to a description opposite to that used here. As a separate but related point, if a curved sheet has a right-handed twist (as described at the beginning of this footnote) when viewed along the direction of the strands, then the twist of the sheet is opposite, i.e., left-handed, when one looks at the sheet perpendicularly to the direction of the chains (see Figure 4 of ref 18).

(16) IUPAC–IUB Commission on Biochemical Nomenclature *Biochemistry* **1970**, *9*, 3471.

(17) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95.

(18) Weatherford, D. W.; Salemme, F. R. *Proc. Natl. Acad. Sci. U.S.A.* **1979**, *76*, 19.

(19) Miller, M. H.; Nêmethy, G.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 470.

(20) It is important to remark, in connection with eq 1, that the graphic approach introduces a coarse graining into the distribution and therefore into any average quantity $\bar{D}$ calculated from eq 1; of course, such coarse graining is not present in the actual distribution. For example, all points in the (κ, τ) plane which fall in a given square of the graph will be considered to be located at the center of that square for the purpose of calculating distances. Thus, two points which are very close together, but which fall in adjacent squares of the graph, are regarded as being 0.1 rad/Å apart. The use of a finer grid for the graph can reduce the effect of this approximation, but a compromise must be struck between the need to sharpen the graph and the inherent limit in the distinctness of points due to experimental error in the X-ray coordinates from which the distribution is constructed. It seems that the (0.1 × 0.1) rad/Å grid represents a good compromise between these two effects.

(21) Equation 2 is adequate for measuring distances within localized structures such as peaks in Figure 2. In general, the two discontinuities discussed in the text must be taken into account when calculating distances between two points in the (κ, τ) plane, and this is done where appropriate throughout this work.

(22) This is in contrast to previous work[23,24] in which we *excluded* α_R helices from the structures identified as bends. The present work demonstrates the structural identity which actually obtains.

(23) Lewis, P. N.; Momany, F. A.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* 1971, *68*, 2293.

(24) Isogai, Y.; Nêmethy, G.; Rackovsky, S.; Leach, S. J.; Scheraga, H. A. *Biopolymers* **1980**, *19*, 1183.

(25) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 291.

(26) Crippen, G. M. *J. Mol. Biol.* **1978**, *126*, 315.

(27) There are only 3729 nearest-neighbor pairs in the 22 proteins because ribonuclease S has one break in the backbone and α-chymotrypsin has two breaks.