Differential Geometry and Polymer Conformation. 4. Conformational and Nucleation Properties of Individual Amino Acids[1]

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

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.
Received March 3, 1982

ABSTRACT: The differential-geometric representation is used to analyze the structures of four-C^α units containing each of the 20 amino acids at the second or third position. With the aid of newly developed methods for the analysis of complete distributions, it is shown that these amino acids can be divided into two groups. On the basis of reasonable hypotheses as to the relationship between (κ,τ) distributions and nucleation on the four-C^α scale, group I, consisting of 13 amino acids, is shown to be responsible for the nucleation of A structure (helices and bends) in both the second and third positions. Group II, which is made up of Pro, Gly, His, Tyr, Cys, Asn, and Trp and which is much smaller than group I in terms of frequency of occurrence, is responsible for nucleation of E (extended) structures. By comparing the frequency of occurrence of the two groups, it is shown that nucleation of A structures predominates over that of E structures on the *four-C^α* length scale.

I. Introduction

In previous papers in this series, we have outlined the development of a differential-geometric (DG) representation of protein backbone structure and illustrated a number of applications thereof. In papers 1[3] and 2,[4] it was shown that the DG representation can serve as the basis of a method for comparing *local* folding of backbone structures in a quantitative manner. In paper 3,[5] the DG approach was used to investigate the early stages of protein folding and to suggest which structures are likely to be formed as folding progresses through the initial steps.

The novel feature of the DG representation which makes these results possible is the fact that it operates on a four-C^α length scale. It therefore makes possible the observation of structural features which are not clearly brought out by other representations—e.g., the (ϕ,ψ) representation, which operates on the single-residue length scale.

In the present work, we shall extend the results of paper 3 by asking *which residues* are likely to be responsible for initiating the various types of structures found in protein backbones on the four-C^α length scale. An alternative way of phrasing this question is to ask what types of structures on the four-C^α scale are induced by each amino acid. The answer will clearly depend on the position of the amino acid in question in the four-C^α unit, since its interactions with residues which are "forward" (i.e., in the C-terminal direction) relative to itself are likely to differ from those with residues which are "backward" (i.e., in the N-terminal direction). We therefore expect that a given amino acid will have a different effect on forward and backward structures. We shall address this point in the following sections.

It is clear that the structure of a four-C^α unit is actually determined by *all* the amino acids contained therein. We are interested, however, in determining what effects are attributable to the presence of a given, *single* amino acid. We shall therefore use data which are averaged over all nearest neighbors.

Before proceeding, we shall sumarize our views on nucleation, which were stated first in paper 3.[5] We consider folding nuclei to be relatively long-lived structures which form in the denatured chain as renaturing conditions are imposed on the system. It is not necessary for such structures to increase the compactness of the chain; one can conceive of situations in which it is actually necessary to keep two parts of the molecule apart in order to guarantee correct folding. Nor is it necessary for such nuclei to be *kinetically* visible in order for them to govern the folding pathway. For example, it is possible that the presence of several very rapidly formed nuclei is necessary in order for the first *slow* folding step to lead to an energetically favorable intermediate structure.

As in paper 3, we shall proceed from a reasonable assumption. Those structures that occur with greater-than-average frequency in native protein structures will be assumed to be of high stability and therefore to have a high probability of forming in the refolding chain. Our approach will therefore be to analyze a sample of protein X-ray structures and construct a distribution of the four-C^α structures associated with each of the 20 naturally occurring amino acids. In section II we discuss the methodology used to generate and analyze these distributions. In section III, we present the results and discuss them, and in section IV we summarize the results.

II. Methods

The protein sample used to generate the individual residue distributions is the same as that used in paper 3.[5] The criteria used to select the proteins are discussed there.

We begin by recalling the basic features of the DG representation. The structure of the N-residue virtual-bond backbone is represented by a set of parameter pairs $\{\kappa_i,\tau_i\}$ $(i = 2, 3, ..., N-2)$. Together, the curvature (κ_i) and torsion (τ_i) at residue i describe the conformation of the four-C^α unit composed of C^α_{i-1} to C^α_{i+2}. The method for constructing $\{\kappa_i,\tau_i\}$ from X-ray coordinates of C^α atoms was given in paper 1.[3]

A convenient method for exhibiting distributions of (κ,τ) values was demonstrated in previous papers.[3-5] This consists of showing the number of occurrences of (κ,τ) within each of a set of (0.1×0.1) rad/Å squares in the (κ,τ) plane. The resulting distribution is a matrix, whose elements are occupation numbers. This type of distribution will constitute the basic data in the present work.

Before proceeding, we establish some notation. We shall be interested in the distributions of (κ,τ) for four-C^α units in which a given amino acid is either at the second or third position (Figure 1). From what was said above, it is clear that, if the given residue is considered to be at position i, these are distributions of (κ_i,τ_i) and $(\kappa_{i-1},\tau_{i-1})$, respectively [since the (κ,τ) characterizing a given four-C^α unit is always assigned to the residue at the second position of the unit]. We shall denote the distribution of (κ_j,τ_j) $(j = i, i-1)$

0024-9297/82/2215-1340$01.25/0

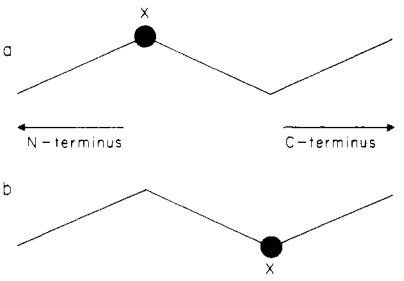

Figure 1. A four-C^α unit, with amino acid X located at the second (a) or third (b) position.

characteristic of a given amino acid X by $\mathbf{X}_j$, e.g., $\mathbf{Ala}_i$ and $\mathbf{Tyr}_{i-1}$. The total (κ_j,τ_j) distribution, irrespective of composition (which is the sum of the individual amino acid distributions), will be denoted by $\mathbf{G}_j$.

We now turn to methods of characterizing distributions. One method, which we shall use for quantitatively characterizing properties within a given *region* of a distribution (consisting of a specified group of squares), is the local number average. If the value of a given property in square (l,m) is given by Z_{lm}, then the average value in a region R is given by

$$\langle Z\rangle_R = \sum_{l,m\subset R} Z_{lm}n_{lm} \Big/ \sum_{l,m\subset R} n_{lm} \tag{1}$$

where n_{lm} is the value of the distribution in square (l,m) and "$\subset$" means "included in".

We shall also be interested in having a method for quantitatively comparing *complete* distributions with each other. Let $\mathbf{X}$ and $\mathbf{Y}$ be the distributions (matrices) that we wish to compare. Consider the quantity

$$\sigma(\mathbf{X},\mathbf{Y}) = \frac{\sum_{l,m} X_{lm}Y_{lm}}{[\sum_{l,m} X_{lm}{}^2]^{1/2}[\sum_{l,m} Y_{lm}{}^2]^{1/2}} \tag{2}$$

where lm denotes elements of the matrix [corresponding to intervals in (κ,τ)]. If we imagine the elements X_{lm} and Y_{lm} to be the components of a vector, then $\sigma(\mathbf{X},\mathbf{Y})$ is the scalar product of two unit vectors, one parallel to $\mathbf{X}$ and one parallel to $\mathbf{Y}$, in a multidimensional space. (The dimensionality of this space, q, is conveniently taken to be the number of nonzero elements of the *total* distribution $\mathbf{G}$; i.e., the dimensionality is the number of occupied squares for *all* amino acids.) The angle between these unit vectors, which is a measure of the distance (or difference) between the distributions, is then given by

$$\theta(\mathbf{X},\mathbf{Y}) = \cos^{-1}[\sigma(\mathbf{X},\mathbf{Y})] \tag{3}$$

It is readily verified that, if the two distributions are similar, i.e., if

$$\mathbf{X} = a\mathbf{Y} \tag{4}$$

where a is a constant, then the vectors are parallel, $\sigma(\mathbf{X},\mathbf{Y}) = 1$ and $\theta(\mathbf{X},\mathbf{Y}) = 0$. This guarantees that $\theta(\mathbf{X},\mathbf{Y})$ is independent of a; i.e., it is not sensitive to the *total* populations represented by the two distributions, but only to the *fractional* distributions—i.e., to the relative disposition of the points represented by the distribution among the various regions of the (κ,τ) plane. It is also clear that, if $\mathbf{X}$ and $\mathbf{Y}$ are completely different [i.e., if they occupy different sets of (κ,τ) squares or, equivalently, if $\mathbf{X}_{lm} \neq 0$ implies $\mathbf{Y}_{lm} = 0$)], then $\theta(\mathbf{X},\mathbf{Y}) = 90°$.[6,7]

As a further useful tool, we define the quantity

$$P_{lm}{}^{(\mathrm{X})} = (X_{lm}/\sum_{l,m} X_{lm})/(G_{lm}/\sum_{l,m} G_{lm}) \tag{5}$$

This quantity is the ratio of the fraction of a given amino acid distribution which falls in square (l,m) to the corresponding fraction for the total distribution irrespective of sequence. If $P_{lm}{}^{(\mathrm{X})} > 1$, the conformation of four-C^α units containing X has a stronger tendency to fall in the region represented by square (l,m) than those of all four-C^α units, irrespective of composition. If $P_{lm}{}^{(\mathrm{X})} < 1$, the opposite is the case. Then $\langle P^{\mathrm{X}}\rangle_R$ (eq 1) gives the average relative tendency of residue X to occur in region R.

In keeping with the hypothesis stated in the Introduction, we suggest that, in the *earliest* stages of folding, four-C^α structures in a given region R are most likely to be nucleated at those residues whose value of $\langle P^{\mathrm{X}}\rangle_R$ is highest—i.e., those for which

$$\langle P^{\mathrm{X}}\rangle_R > \langle \bar{P}\rangle_R + \sigma_R \tag{6}$$

where σ_R is the standard deviation of the set of $\langle P^{\mathrm{X}}\rangle_R$ and $\langle \bar{P}\rangle_R$ is the mean. (Here we have chosen σ_R, viz., one standard deviation, as a somewhat arbitrary but nevertheless reasonable criterion for differentiating between conformational tendencies.) As conditions in later stages become more favorable for renaturation, structure in region R is likely to form at residues characterized by $\langle P^{\mathrm{X}}\rangle_R$ such that

$$\langle \bar{P}\rangle_R < \langle P^{\mathrm{X}}\rangle_R < \langle \bar{P}\rangle_R + \sigma_R \tag{7}$$

A given structure is likely to form at residues for which

$$\langle \bar{P}\rangle_R - \sigma_R < \langle P^{\mathrm{X}}\rangle_R < \langle \bar{P}\rangle_R \tag{8}$$

at a still later stage of folding, probably under the influence of interactions that act on a length scale larger than that of a four-C^α unit. In other words, inequalities 6–8 describe the conditions that residue X induces the formation of four-C^α structure in region R in the earliest, somewhat later, and still later stages of folding. Structures for which

$$\langle P^{\mathrm{X}}\rangle_R < \langle \bar{P}\rangle_R - \sigma_R \tag{9}$$

are assumed to be strongly avoided and to form only under the influence of long-range interactions.

III. Results and Discussion

There are six principal structural regions in the (κ,τ) plane[5]—three helix/bend regions, A_R, A_0, and A_L, and three extended regions, E_R, E_0, and E_L. Here R, L, and 0 denote right-handed, left-handed, and nearly flat four-C^α structures. (The reader is referred to paper 3[5] for the definitions of these regions and a discussion of their significance.) The values of $\langle P^{\mathrm{X}}\rangle_R$ were generated for each of the 20 amino acids for all six of these regions and are shown in Table I. Data are shown for both $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$.

The data in Table I are sufficient to indicate that there are differences in conformational properties between the various amino acids on the four-C^α length scale. This observation raises a number of questions, which we shall address in this work:

(1) Which amino acids exhibit extremes of conformational behavior—either great differences from or strong similarity to that of other amino acids?
(2) How can this global information be correlated with the data in Table I on the behavior of the amino acids in specific conformational regions?
(3) Is there any correlation between the conformational properties of amino acids and the frequency with which they occur in the protein sample?

The last question is particularly intriguing, since a positive answer may reveal a disparity in the frequency of formation of various nucleation structures—or, alternatively stated, a conformational reason for the wide variation in the frequency of occurrence of the various

Table I
Average Relative Fractional Occurrence $\langle P^X \rangle_R$ in the Various Structural Regions R

	X	A_0	A_R	A_L	E_L	E_0	E_R
				A. (κ_i,τ_i) Distributions			
1	Ala	0.85	1.58	0.82	0.78	0.88	0.30
2	Asp	1.50	0.98	2.64	1.25	1.16	1.26
3	Cys	0.90	1.04	0.00	3.14	1.14	0.72
4	Glu	1.79	1.49	2.62	0.94	1.01	1.33
5	Phe	0.85	1.00	0.00	1.07	1.52	1.20
6	Gly	1.54	0.66	1.63	1.13	0.70	3.09
7	His	1.59	0.99	0.00	1.03	1.87	1.33
8	Ile	0.67	1.09	2.32	1.26	1.61	0.45
9	Lys	0.88	1.27	2.86	0.85	0.83	0.71
10	Leu	1.03	1.21	0.00	0.91	1.09	0.96
11	Met	1.17	1.41	0.00	0.41	1.71	1.89
12	Asn	0.88	0.77	2.07	1.32	1.02	2.73
13	Pro	1.47	1.46	0.00	1.73	0.87	0.83
14	Gln	1.71	1.24	0.00	0.93	0.93	0.97
15	Arg	2.02	1.14	2.60	1.75	0.99	0.90
16	Ser	1.50	1.05	1.23	1.31	1.14	1.16
17	Thr	1.96	0.87	2.48	1.57	0.96	0.97
18	Val	0.89	0.88	1.62	1.11	1.56	0.64
19	Trp	0.83	1.23	0.00	0.98	1.96	1.58
20	Tyr	1.34	0.68	1.90	1.31	1.68	0.86
				B. $(\kappa_{i-1},\tau_{i-1})$ Distributions			
1	Ala	0.40	1.48	0.00	1.02	0.93	0.99
2	Asp	1.59	1.19	2.15	1.76	0.60	1.18
3	Cys	2.98	0.86	0.00	1.05	1.08	2.32
4	Glu	1.26	1.43	0.00	0.83	0.73	1.36
5	Phe	1.27	1.30	2.11	0.41	1.51	1.25
6	Gly	1.89	0.46	6.49	2.39	0.78	1.40
7	His	2.71	1.27	0.00	0.40	1.08	1.06
8	Ile	1.31	1.12	0.00	0.83	1.74	0.81
9	Lys	0.87	1.36	0.00	0.94	1.00	0.91
10	Leu	0.57	1.33	0.00	1.06	1.03	1.26
11	Met	0.00	1.41	0.00	1.33	1.31	1.00
12	Asn	1.24	0.99	4.14	1.31	0.92	1.15
13	Pro	0.38	0.25	1.99	2.73	1.37	0.00
14	Gln	0.50	1.42	0.00	1.05	0.94	1.52
15	Arg	1.20	1.02	0.00	1.00	1.52	1.19
16	Ser	0.92	0.89	0.00	1.18	0.97	1.50
17	Thr	1.38	0.81	1.24	0.77	1.38	1.18
18	Val	0.95	0.93	0.00	0.88	1.70	1.01
19	Trp	1.53	1.27	0.00	1.22	1.12	1.33
20	Tyr	1.79	0.91	1.90	1.09	1.65	1.09

amino acids. We shall address the three questions in turn.

A. Question 1. Recall that we represent the $\mathbf{X}_j$ (j = i, i–1) as unit vectors in a q-dimensional space. A natural axis in this q-dimensional space, relative to which we can measure the orientations of the $\mathbf{X}_j$, is formed by the unit vector corresponding to $\mathbf{G}_j$, the total distribution. We therefore calculate values of $\theta(\mathbf{G}_i,\mathbf{X}_i)$ and $\theta(\mathbf{G}_{i-1},\mathbf{X}_{i-1})$ (Table II). These values of θ can be thought of as giving the distribution of *polar* angles of $\mathbf{X}_j$ with $\mathbf{G}_j$ (by analogy with a three-dimensional spherical coordinate system) and therefore provide a partial answer to question 1 above, by showing which amino acids have (κ,τ) distributions significantly different from the total distribution. (A value of θ = 0 means that the residue in question has a distribution which is identical with the total distribution.) In order to complete our answer, it will be necessary in effect to investigate the *azimuthal* distribution of the $\mathbf{X}_j$ about $\mathbf{G}_j$ in the q-dimensional space. We shall presently do this by calculating $\theta(\mathbf{X}_i,\mathbf{Y}_i)$ and $\theta(\mathbf{X}_{i-1},\mathbf{Y}_{i-1})$.

It will be seen from Table II that the polar distributions of both $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$ are, for the most part, fairly narrow. The average value of $\theta(\mathbf{G}_j,\mathbf{X}_j)$, which we denote by $\bar{\theta}_j$, is ~22° in each case. In both cases, furthermore, $\theta(\mathbf{G}_j,\mathbf{X}_j) < \bar{\theta}_j$ for a majority of amino acids, and, in fact a majority of amino acids have $\theta(\mathbf{G}_j,\mathbf{X}_j)$ which fall in an interval of width ~12°. (It should be noted that these differences

Table II
Values of $\theta(\mathbf{G}_i,\mathbf{X}_i)$ and $\theta(\mathbf{G}_{i-1},\mathbf{X}_{i-1})$ in Degrees

$\theta(\mathbf{G}_i,\mathbf{X}_i)$		$\theta(\mathbf{G}_{i-1},\mathbf{X}_{i-1})$	
X	θ	X	θ
Leu	10.89	Asn	13.59
Gln	15.42	Lys	14.07
Phe	16.26	Leu	14.30
Lys	16.46	Ala	14.53
Ile	16.66	Arg	17.82
Ala	17.05	Glu	18.19
Val	17.06	Ser	18.56
Thr	18.92	Phe	19.61
Asp	19.27	Asp	19.78
Ser	19.95	Trp	19.78
Glu	20.12	Ile	20.28
Met	20.61	Met	20.61
Arg	21.25	Thr	21.09
His	23.07	Val	21.87
Trp	23.36	Gln	22.18
Pro	23.94	His	22.63
Tyr	26.49	Tyr	26.36
Cys	28.84	Cys	30.57
Asn	34.81	Gly	37.16
Gly	38.14	Pro	52.63
mean	21.43 ± 6.6	mean	22.28 ± 9.1

are statistically significant. Use of the χ^2 test shows that all the $\mathbf{X}_j$ differ from $\mathbf{G}_j$ at a confidence level greater than

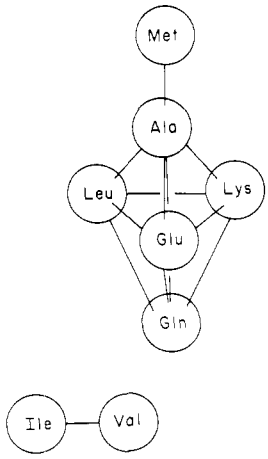

Figure 2. Amino acids connected by lines are those with very similar (κ_i,τ_i) and $(\kappa_{i-1},\tau_{i-1})$ distributions.

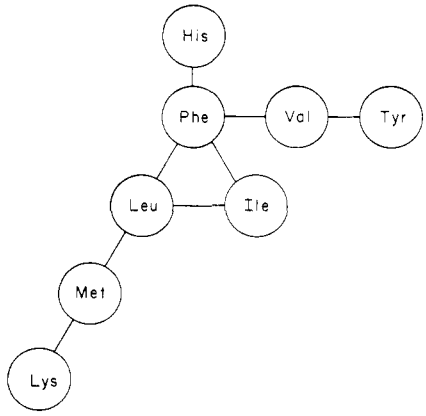

Figure 3. Amino acids connected by lines are those with very similar (κ_i,τ_i) distributions only.

99%.) There are, however, a number of amino acids whose distributions are *very* different from the total distribution. Inspection of Table II shows that five amino acids—Pro, Gly, His, Tyr, and Cys—have $\theta(\mathbf{X}_j,\mathbf{G}_j) > \bar{\theta}_j$ for *both* $j = i$ and $j = i - 1$. Two others, Asn and Trp, have only $\theta(\mathbf{X}_i,\mathbf{G}_i) > \bar{\theta}_i$. Asn, in fact, has the peculiar property that its conformational behavior in the *forward* direction (i.e., that of $\mathbf{Asn}_i$) is *very* different from that of $\mathbf{G}_i$, whereas in the backward direction (represented by $\mathbf{Asn}_{i-1}$) it is the most "normal" of amino acids (i.e., most similar to $\mathbf{G}_{i-1}$). Trp, on the other hand, is at nearly the average distance from $\mathbf{G}_j$ in both cases.

That these amino acids indeed differ in their conformational properties is confirmed by calculating $\theta(\mathbf{X}_i,\mathbf{Y}_i)$ and $\theta(\mathbf{X}_{i-1},\mathbf{Y}_{i-1})$ (Tables III and IV). It can be seen that certain amino acids consistently show values of $\theta(\mathbf{X}_j,\mathbf{Y}_j) > \hat{\theta}_j$ (where $\hat{\theta}_j$ is the mean value of $\theta(\mathbf{X}_j,\mathbf{Y}_j)$ for *all* amino acid pairs given in the tables). This fact is summarized in Table V, which shows the number of $\theta(\mathbf{X}_j,\mathbf{Y}_j) > \hat{\theta}_j$ for each amino acid. These values correlate precisely with the observation made above on the basis of the $\theta(\mathbf{G}_j,\mathbf{X}_j)$. The same group of amino acids—Pro, Gly, His, Tyr, Cys, Asn, and Trp—exhibit conformational behavior on the four-C^α length scale, either in the forward or backward direction or both, which is significantly different from that of the remaining amino acids.

This observation answers one part of question 1 above. The answer to the second part—the identification of pairs of amino acids with very similar conformational behavior—is also contained in Tables III and IV. The results are summarized in Figures 2–4, in which lines connect those amino acids whose four-C^α conformational

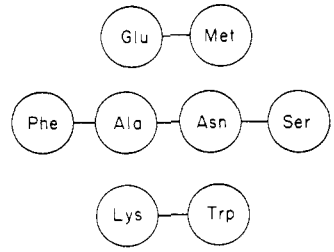

Figure 4. Amino acids connected by lines are those with very similar $(\kappa_{i-1},\tau_{i-1})$ distributions only.

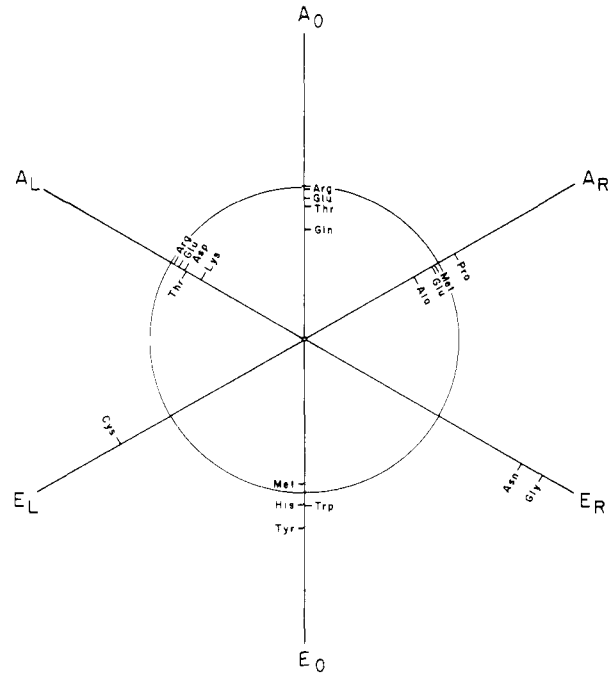

Figure 5. Plot of $\theta(\mathbf{G}_i,\mathbf{X}_i)/\bar{\theta}_i$ *for those amino acids which satisfy eq 6* (i.e., which have a very strong tendency to occur in the given region) for each of the six structural regions in the (κ,τ) plane. $\theta(\mathbf{G}_i,\mathbf{X}_i)/\bar{\theta}_i$ increases from 0 radially along each axis. The circle indicates the radius at which $\theta(\mathbf{G}_i,\mathbf{X}_i)/\bar{\theta}_i = 1.0$. Those amino acids with $\theta(\mathbf{G}_i,\mathbf{X}_i)/\bar{\theta}_i > 1.0$ (i.e., those which fall outside the circle) have (κ_i,τ_i) distributions which differ from the total distribution $\mathbf{G}_i$ by more than the average amount. Values of $\theta(\mathbf{G}_i,\mathbf{X}_i)$ and $\bar{\theta}_i$ are from Table II. It should be noted that the structural region corresponding to each of the six axes is adjacent, in the (κ,τ) plane, to those corresponding to its two neighboring axes in the figure.

behavior is very similar [as defined by $\theta(\mathbf{X}_j,\mathbf{Y}_j) < \hat{\theta}_j - \sigma$, where σ is the standard deviation of $\theta(\mathbf{X}_j,\mathbf{Y}_j)$]. In Figure 2 we show those amino acids with similar behavior for both $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$. Figure 3 shows those with similar behavior for $\mathbf{X}_i$ only, and Figure 4 shows those with similar behavior for $\mathbf{X}_{i-1}$ only. On the basis of four-C^α structure, it is to be expected that substitutions between amino acids which are connected in Figure 2 will occur very readily, for example in variable regions of homologous proteins, with those indicated in Figures 3 and 4 occurring with rather lower readiness. It should be noted, however, that the side chains of these residues vary considerably in structure, polarity, etc. Therefore long-range interactions (hydrogen bonds and Coulomb interactions) undoubtedly rule out some substitutions which might have been acceptable on the four-C^α length scale.

B. Question 2. Having determined which amino acids show extremes of behavior, we shall determine what specific conformational characteristics cause this behavior and thereby answer question 2 above. In other words, we wish to correlate the global and local characteristics of the (κ,τ) distributions. We do this by noting which amino acids

Table III
Values of $\theta(X_i, Y_i)$ (Deg)[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 Ala		25.86	36.79	14.54	25.05	48.78	29.94	24.90	12.90	15.25	21.65	45.03	32.68	17.55	24.71	33.00	28.52	28.57	27.56	38.70
2 Asp			38.00	26.36	27.57	39.29	25.76	31.35	27.69	23.81	28.80	32.62	32.18	25.32	29.06	25.92	28.78	28.29	31.17	35.90
3 Cys				38.62	30.82	47.65	37.01	31.19	34.45	33.23	38.54	42.07	36.20	37.97	32.46	30.71	30.75	30.84	40.82	32.29
4 Glu					28.67	48.83	32.04	28.27	16.28	18.48	24.42	46.79	32.18	18.03	25.94	35.31	30.53	31.35	33.71	42.35
5 Phe						42.88	22.36	20.23	25.94	21.68	24.30	38.84	28.79	23.62	27.62	28.76	24.52	17.87	28.56	21.52
6 Gly							44.54	45.32	45.30	41.75	49.87	35.02	45.72	45.89	47.46	36.57	41.93	45.68	47.71	44.72
7 His								27.65	31.86	27.97	33.09	40.32	32.60	27.27	33.23	32.36	29.52	26.96	32.66	30.99
8 Ile									23.51	16.80	23.98	45.76	25.92	20.61	28.76	26.35	26.57	16.59	25.39	28.35
9 Lys										15.05	22.54	45.80	31.04	18.55	24.95	29.25	27.29	29.12	29.29	38.85
10 Leu											18.87	41.70	26.43	15.64	25.98	26.23	25.95	21.33	23.00	32.27
11 Met												45.11	37.23	22.35	31.47	31.03	33.35	22.42	21.81	34.81
12 Asn													44.07	44.21	42.79	33.59	33.28	38.80	38.83	37.64
13 Pro														25.64	30.55	27.92	27.18	29.09	38.95	34.76
14 Gln															25.81	28.88	27.12	23.81	28.04	35.15
15 Arg																30.15	27.07	27.96	36.48	32.14
16 Ser																	26.29	24.57	29.37	30.42
17 Thr																		25.12	33.88	27.12
18 Val																			23.15	19.67
19 Trp																				34.34
20 Tyr																				

[a] Mean: $\hat{\theta} = 30.90°$; standard deviation = 8.09°.

Table IV
Values of $\theta(X_{i-1}, Y_{i-1})$ (Deg)[a]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1 Ala		25.14	38.58	12.91	20.81	46.24	23.68	26.55	15.36	15.28	18.08	20.48	60.60	20.64	22.78	24.85	31.21	30.57	22.71	34.17
2 Asp			35.51	24.18	31.77	39.32	30.91	32.42	21.74	23.93	27.33	22.35	54.82	23.75	26.42	27.16	33.64	34.82	26.97	36.94
3 Cys				40.31	37.69	40.09	33.32	33.84	36.64	35.95	43.00	34.05	57.06	38.85	37.37	33.35	32.17	34.75	37.26	40.32
4 Glu					22.40	44.48	24.34	33.06	18.62	20.34	20.04	21.30	64.75	20.31	22.35	26.75	34.46	36.54	24.67	34.53
5 Phe						46.49	23.55	26.95	22.13	25.46	30.84	24.68	65.03	31.72	27.43	27.68	25.79	28.62	23.55	30.10
6 Gly							46.58	45.03	44.02	45.95	47.98	36.24	50.63	46.80	42.37	38.81	42.77	45.05	44.48	40.85
7 His								27.14	24.19	24.21	32.07	30.54	69.10	26.32	29.74	34.65	29.55	30.94	27.73	35.68
8 Ile									26.60	24.24	33.10	29.42	55.58	32.50	30.22	32.88	23.00	13.59	32.08	31.45
9 Lys										12.71	23.59	22.19	59.35	20.93	22.91	28.13	27.81	27.97	19.10	35.75
10 Leu											22.10	23.72	58.35	18.06	23.02	28.55	27.14	26.30	22.77	33.98
11 Met												21.44	56.76	25.24	24.12	24.00	35.15	34.97	28.71	34.82
12 Asn													49.26	28.08	22.08	15.50	27.64	30.13	23.53	27.29
13 Pro														64.19	56.27	47.88	53.78	51.32	60.44	53.79
14 Gln															29.10	32.55	36.52	35.57	29.85	41.97
15 Arg																22.66	26.83	31.49	24.26	28.72
16 Ser																	30.05	32.88	27.44	27.01
17 Thr																		18.94	25.96	29.14
18 Val																			31.89	33.03
19 Trp																				35.68
20 Tyr																				

[a] Mean: $\hat{\theta} = 32.40°$; standard deviation = 11.16°.

Table V
Number of Values of $\theta(X_j,Y_j) > \hat{\theta}_j$

$j = i$		$j = i-1$	
Phe	2	Phe, Asn, Arg	3
Val	3	Ala, Lys, Leu, Trp	4
Ile, Gln, Thr	4	His	5
Ala, Lys, Leu, Ser	6	Thr	6
Asp, Arg	7	Asp, Glu, Met, Ser	7
Glu, Met	9	Ile, Gln	8
Trp	10	Val	9
His, Pro	11	Tyr	13
Tyr	14	Cys	18
Cys	15	Gly, Pro	19
Gly, Asn	19		
	8.6 ± 4.98		7.9 ± 5.25

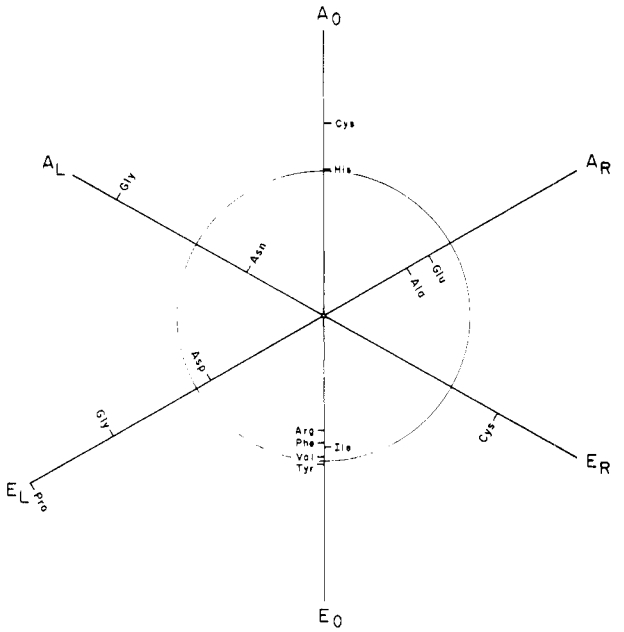

Figure 6. Same plot as in Figure 5 for $\theta(\mathbf{G}_{i-1},\mathbf{X}_{i-1})/\bar{\theta}_{i-1}$. Values of $\theta(\mathbf{G}_{i-1},\mathbf{X}_{i-1})$ and $\bar{\theta}_{i-1}$ are from Table II.

show a strong tendency to induce a given type of structure in four-C^α units in which they occur, as defined by eq 6. This information is shown in Figures 5 and 6, correlated with the values of $\theta(\mathbf{G}_j,\mathbf{X}_j)$ for each amino acid. This plot brings out very clearly the conformational basis of the distinction between the two groups of amino acids which we identified previously using values of $\theta(\mathbf{X}_j,\mathbf{Y}_j)$ and $\theta(\mathbf{G}_j,\mathbf{X}_j)$. In the forward direction, as shown in Figure 5, essentially all induction of A structure—A_R helices and A_R, A_0, and A_L bends—is by members of the larger group of amino acids with $\theta(\mathbf{G}_i,\mathbf{X}_i) \lesssim \bar{\theta}_i$ (which we shall refer to henceforth as group I). Almost all induction of E (extended) structure is by members of the smaller group (group II, consisting of Pro, Gly, His, Tyr, Cys, Asn, and Trp). The role of group II amino acids is even more significant in the induction of structure in the backward direction (Figure 6), where they are essentially responsible for the induction of all structures except E_0 and A_R.

C. Question 3. In view of the clear distinction between the nucleation behavior of the two groups, we can obtain

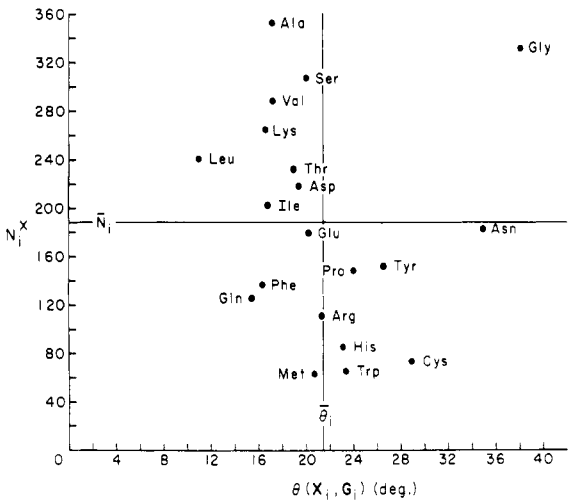

Figure 7. Plot of N_i^X vs. $\theta(\mathbf{X}_i,\mathbf{G}_i)$ (see text). Values of $\bar{N}_i$, the average value of N_i^X, and $\bar{\theta}_i$ are shown.

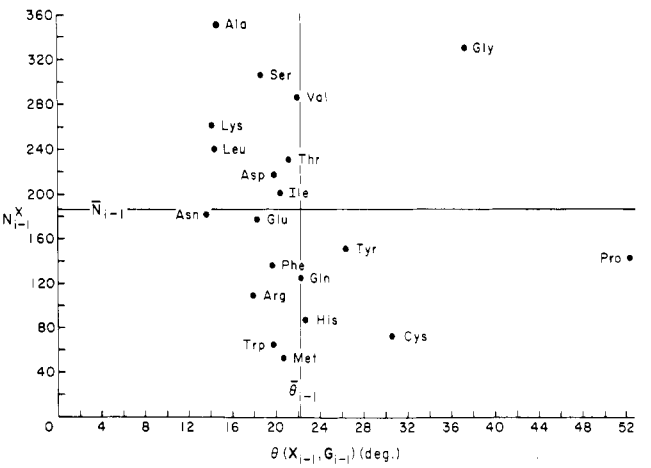

Figure 8. Same as Figure 7 for N_{i-1}^X and $\theta(\mathbf{X}_{i-1},\mathbf{G}_{i-1})$.

information as to the relative frequency of nucleation of different types of structure by correlating the conformational behavior of the various amino acids with their frequencies of occurrence in the protein sample on which the data are based. This correlation is shown in Figures 7 and 8, in which N_j^X, the number of occurrences of X at the appropriate (second or third) position of four-C^α units, is plotted against $\theta(\mathbf{G}_j,\mathbf{X}_j)$. It can be seen that, with the exception of Gly, *all* the members of group II occur with less-than-average frequency. This suggests strongly that the *principal* nucleation events on the four-C^α length scale will be those governed by group I amino acids, viz., the formation of A structures, under the influence of either the second or (in the case of A_R) the third amino acid in the four-C^α unit. The formation of E structures will be much less common, with the formation of E_0 expected to dominate among these (since, of all the E structures, E_0 is induced by the largest number of amino acids, in either the second or third position).

We have thus answered the third question posed above. It seems correct to conclude that the nucleation of E structures is assigned to the less-common amino acids in order to guarantee that most (although not all) nucleation steps increase the compactness of the structure through the formation of A structures. It should be noted, though, that not all the less-common amino acids fall into group II. In fact, group II amino acids account for only 28% of the residues in the sample. It is possible that the other less-common amino acids have special conformational properties on a different length scale.

The structural basis for the unusual behavior of the group II amino acids is necessarily different in each case and is best elucidated through detailed conformational energy calculations, e.g., those of Zimmerman et al.[8,9] Nevertheless, certain unusual characteristics of the group II amino acids exist, which might possibly contribute to their four-C^α conformational behavior. For example, Pro is known to cause formation of unusual structures immediately N-terminal to itself (e.g., cis peptide bonds) because of its unique imino acid structure.[9] Gly, which lacks a side chain, has greater conformational freedom than any other amino acid. The conformational properties of Cys are necessarily governed by the constraints imposed by the formation of disulfide bonds. The very interesting behavior of Asn is probably related to the relatively high tendency of this residue to fall in the α_L region of the (ϕ,ψ) map.[10] [Nevertheless, we again emphasize the fact that there is only a very general correlation between single-residue (ϕ,ψ) behavior and (κ,τ) behavior, since the latter is governed by *two* sets of (ϕ,ψ).] The behavior of His, Trp, and Tyr is less dramatic and also less readily correlated with structure.

The current work takes a somewhat different approach to correlating composition and structure than previous work, in which the amino acid composition of specific structural features, e.g., bends,[11] was studied. Here we study the structural features on the four-C^α length scale associated with particular amino acids. The resulting (κ,τ) maps can be regarded as empirical four-C^α conformational energy maps, averaged over nearest neighbors, in the same sense that individual residue empirical (ϕ,ψ) maps are known to correspond closely to calculated potential energy maps.

IV. Summary

We have examined the (κ_i,τ_i) and $(\kappa_{i-1},\tau_{i-1})$ distributions associated with each of the 20 naturally occurring amino acids. Analysis of the resulting maps of structure of four-C^α units reveals a number of features.

(1) There is a group of amino acids (group II, consisting of Pro, Gly, His, Tyr, Cys, Asn, and Trp) which show significantly different conformational behavior than the majority, when located at either the second or third position of a four-C^α unit.

(2) Pairs of amino acids can also be identified which show very similar conformational behavior on the four-C^α length scale when located at the same position in the four-C^α unit.

(3) On the basis of reasonable postulates as to the relationship between the statistical properties of the (κ,τ) maps and the nucleation properties of the amino acids, it is demonstrated that group I and group II amino acids have distinctly different nucleating properties. Group I residues induce the formation of A structures—A_R helices and A_R, A_L, and A_0 bends— in four-C^α units in which they are located at the second position, and E_0 and A_R structures when located at the third position. Group II residues are responsible for nucleation of E structures in four-C^α units in which they are at the second position, and all structures except E_0 and A_R when at the third position.

(4) By relating these observations to the composition of the protein data base, it is shown that the formation of A structures must be the dominant nucleation process, although E structures are also formed.

This view of nucleation is broad enough to encompass various proposed nucleation mechanisms. Both the hairpin bend proposed by Matheson and Scheraga[12] and the α helix proposed by Ptitsyn[13] as primary nucleating structures have their origin in A structures. Apparently the distinction between these mechanisms will have to be made on a longer length scale than the four-C^α unit. We reiterate, however, that the nucleation structures to which we refer are possibly, but *not necessarily*, the first kinetically visible structure. They may also be rapidly formed structures whose presence is necessary to ensure that the first kinetically visible folding steps lead to energetically favorable structural intermediates.

Acknowledgment. We thank Dr. George Némethy for helpful conversations and comments on the manuscript.

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) To whom requests for reprints should be addressed.
(3) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 1168.
(4) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1980**, *13*, 1440.
(5) Rackovsky, S.; Scheraga, H. A. *Macromolecules* **1981**, *14*, 1259.
(6) After this work was completed, we became aware of recent work of Kolaskar and Ramabrahmam[7] comparing (ϕ,ψ) distributions. Their comparison method is different from ours, and does not have all of the mathematical properties of $\theta(\mathbf{X},\mathbf{Y})$. Their results, of course, pertain to the single-residue rather than the four-C^α length scale.
(7) Kolaskar, A. S.; Ramabrahmam, V. *Int. J. Biol. Macromol.* **1981**, *3*, 171.
(8) Zimmerman, S. S.; Pottle, M. S.; Némethy, G.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 1.
(9) Zimmerman, S. S.; Scheraga, H. A. *Biopolymers* **1977**, *16*, 811.
(10) Némethy, G.; Scheraga, H. A. *Q. Rev. Biophys.* **1977**, *10*, 239.
(11) Zimmerman, S. S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 4126.
(12) Matheson, R. R., Jr.; Scheraga, H. A. *Macromolecules* **1978**, *11*, 819.
(13) Ptitsyn, O. B. *FEBS Lett.* **1981**, *131*, 197.