J. Mol. Biol. (1974) 88, 457–469

The Structural Code for Proteins : Zonal Distribution of Amino Acid Residues and Stabilization of Helices by Hydrophobic Triplets

J. PALAU AND P. PUIGDOMÉNECH

Instituto de Investigaciones de Biología Fundamental
Centro Coordinado del C.S.I.C., IBF
Universidad Autónoma de Barcelona
Avda San Antonio Mª Claret, 171. Barcelona-13, Spain

(Received 28 February 1974, and in revised form 18 June 1974)

The general distribution of 2259 amino acid residues in α-helical and in non-helical regions has been calculated using data from 12 proteins. The irregular distribution of hydrophobic residues in the helical parts of the sample permits classification of the helical regions in hydrophobic-clustered and hydrophobic-depleted zones. In comparison with the composition of α-helices, Ala, Leu, Val, Ile and Tyr predominate in hydrophobic-clustered zones, whereas Cys and Gln tend to accumulate in hydrophobic-depleted zones. The N-terminal part of the helices is rich in Asp and Pro and poor in Leu and His, whereas the C-terminal is rich in Lys and Gln and does not contain Pro.

For helical regions, hydrophobic residues, located along the sequence with a relation of 1-2-5 or 1-4-5, are more frequent than those with a relation of 1-2-3, 1-2-4, 1-3-4 or 1-3-5. The opposite is found for non-helical regions. All these deviations are statistically significant. It is concluded that hydrophobic triplets 1-2-5 and 1-4-5 are required for stabilizing the helices.

The relevance of our results, and others reported in the bibliography, towards a better understanding of the structural code for proteins is discussed.

1. Introduction

The X-ray diffraction technique has emerged as the most useful tool for conformation studies of proteins. The high resolution of this technique permits very precise location of the regions of a protein having a definite secondary structure such as α-helix or β-conformation. Therefore, for a number of proteins it is possible to correlate primary structure with structures of higher order.

Current theory of molecular biology establishes that linear genetic information is sufficient to code the whole conformation of a given protein. The structural code for proteins, i.e. the code which allows translation of the primary structure into structures of higher order, is still far from being established. Pain & Robson (1970), Robson & Pain (1971) and Ponnuswammy *et al.* (1973) have outlined three outstanding sequential events in the process of spatial ordering of the amino acid residues in a polypeptide chain: short-range interactions, medium-range interactions and long-range interactions. In a comprehensive way, partial codes should exist for helices, β-sheets, loops (including structures such as reversal turns; Crawford *et al.*, 1973),

and random coils, the last being presumably produced by nonsense informational orders.

Until now, most efforts have been devoted to the elucidation of the code for α-helices. In this sense, different authors have attempted to correlate the amino acid sequence of proteins with the potential for α-helix formation and, accordingly, they have developed several predictive procedures. In short, and after Low *et al.* (1968), we can classify these procedures as follows: (a) on the statistical analysis of the distribution of several amino acids considered as α-helical stabilizers or α-helical destabilizers (Guzzo, 1965; Prothero, 1966; Cook, 1967; Kotelchuck & Scheraga, 1969; Ptitsyn, 1969); (b) on the observation that intrahelical interactions may imply residues located at one side of the helix, such as residues n and $n \pm 3$ or $n \pm 4$, these interactions being stabilizing if the residues are hydrophobic (Perutz *et al.*, 1965; Schiffer & Edmunson, 1967); (c) on calculations of the helical potential for a given residue, taking into account the specific environment of polypeptide sequence (Periti *et al.*, 1967); (d) a predictive method based on comparisons of the primary structure of proteins of known and unknown secondary structure (Low *et al.*, 1968).

During the last few years studies have been focussed on searching helical potentials of individual residues (Pain & Robson, 1970; Finkelstein & Ptitsyn, 1971; Leberman, 1971; Chou & Fasman, 1973; Crawford *et al.*, 1973) and on exploring the influence of medium-range interactions (Kotelchuck *et al.*, 1969; Krigbaum & Rubin, 1971; Robson & Pain, 1971; Wu & Kabat, 1971; Nagano, 1973; Wu & Kabat, 1973) using more refined mathematical methods than before, and taking advantage of the fact that an increasing number of protein structures have been resolved.

In spite of such a great effort, many aspects still remain obscure and results given by different authors are, quite often, controversial. The work described in this paper concerns procedures (a) and (b) above, and it represents an attempt to clarify the following points.

(1) The presence in the α-helical regions of hydrophobic-clustered and hydrophobic-depleted zones.

(2) The α-helical forming and breaking effect, as given by the amino acid distribution of N-terminal and C-terminal triplets in α-helices.

(3) The stabilizing effect of hydrophobic arcs, under the assumption that hydrophobic residues placed in a given order (1-2-5 or 1-4-5) are helical stabilizers.

2. Materials and Methods

Twelve proteins whose primary and secondary structure are known have been the object of our analysis (see Table 1). Only one from each family of proteins was chosen, with the exception of myoglobin and haemoglobin (which are rather similar in sequence and structure). The sequence and location of α-helical segments were taken from recent reviews or general books available, but the original articles were consulted in order to unify criteria. The total number of residues of the sample is 2259, 790 of them being in helical regions, i.e. 35%. As in a related paper (Schiffer & Edmunson, 1967), the following amino acids were considered as hydrophobic: Leu, Ile, Val, Met, Phe, Tyr, Trp and Ala.

Analysis of protein sequences and the main statistical calculations were carried out on an IBM360 computer and programs were written in FORTRAN IV. Amino acid percentages, χ^2 tests and auxiliary calculations were carried out on an Olivetti Programma

101. With the exception of combinatory analyses, statistical calculations correspond to current treatments found in general textbooks of statistics.

Several computational methods have been used.

(a) *Analysis of amino acid composition of different regions and zones of the sample*

(b) *Analysis of hydrophobic triplets in helical and in non-helical regions*

The number and position of different types of triplets of hydrophobic residues (the triplets which appear in a five-residue segment are: 1-2-3, 1-2-4, 1-2-5, 1-3-4, 1-3-5 and 1-4-5) were computed for α-helical and for non-helical regions. Triplets which belong partly to helical and partly to non-helical regions were not considered. The compositions of all triplets 1-2-5 and 1-4-5 were also listed. The theoretical probability of finding hydrophobic triplets in the sample was calculated as follows:

$$q = \frac{N_H}{N_T}\frac{N_H - 1}{N_T - 1}\frac{N_H - 2}{N_T - 2},$$

where N_H is the total number of hydrophobic residues and N_T the total number of residues in the corresponding regions (α-helical or non-helical).

The experimental frequency for each hydrophobic triplet within each region was calculated by dividing the number of each hydrophobic triplet by the number of "computer counts" of the given triplet in a given region.

(c) χ^2 *test for statistical reliability of deviations*

The deviation of the proportions of amino acids between two different sets of values was tested by means of a standard χ^2 test. The χ^2 variable is defined as:

$$\chi^2 = \sum_{r=1}^{20} \frac{(N_{sr} - \frac{N_s}{N_T} N_{Tr})^2}{\frac{N_s}{N_T} N_{Tr}},$$

where subscript r refers to a given residue, T to the total region and s to the subgroup being tested; χ^2 is compared $\chi^2_{\alpha,\,19}$, the number of degrees of freedom being $19 = 20 - 1$, and α the security chosen.

(d) *Other general tests*

The significance of the results of the triplet analysis was studied by two different methods.

(1) The number of each kind of triplet present in each protein was computed. The value of the number of triplets normalized with respect to the number of computer counts was calculated for each possible combination of N proteins, N varying from 1 to 11. Collections of values for the experimental frequency were obtained (66 for combinations of orders 10 or 2, 220 for combinations of orders 9 or 3, etc.). The average value and the standard deviation of these collections were computed. The variation of both values with respect to the order of the combination gives information about the extent of the spreading of statistical samples smaller than 12 proteins.

(2) A collection of 8 proteins out of the 12 was taken at random. Four series of 8 values of experimental frequency may be obtained, each protein in the collection being substituted by the other 4, which are not among the 8 chosen. The 6 correlation coefficients for the 4 series of values were calculated. It is possible in this way to detect the deviations of the results for individual proteins from the global results.

(e) *Tests for statistical reliability of deviations for residues and triplets*

This methodology was used for studying the significance of deviations in the composition of each amino acid residue within a region (or zone) as compared with that expected

from a normal distribution of the residues found in the total sample (or in a given region). For this purpose the normal distribution was made equivalent to a binomial distribution. The variance σ_{rs}^2 for a given residue r within a given subgroup s is:

$$\sigma_{rs}^2 = q_r (1 - q_r) N_s.$$

In this expression, q_r is the probability of finding a residue r, and N_s is the total number of residues within each subgroup considered. The statistical reliability was measured by

$$u_{rs} = \frac{\Delta n_{rs}}{\sigma_{rs}}.$$

Here Δn_{rs} is the deviation found for each amino acid residue. Values higher than 2·576, 1·960 and 1·645 correspond respectively to 99·5% ($\pm\pm\pm$), 97·5% ($\pm\pm$), and 95% ($\pm$) of reliability that the deviations are not simple statistical ones.

A similar treatment was carried out for the calculation of the theoretical standard deviations for the hydrophobic triplets. In this case we have

$$\sigma_{rs}^2 = q (1 - q) N_s.$$

In this expression q is the theoretical probability defined in section (b) above, and N_s is the number of computer counts for each triplet in a given region.

3. Results

(a) *General distribution of residues in helical regions*

For α-helical regions we found that Ala, Leu, His and Glu present large positive deviations, as compared with the total sample, whereas Gly and Pro, and to a lesser extent Ser and Asn, present negative deviations. For non-helical regions, Gly and Pro largely exceed the expected value and Ala and Leu show negative deviations that appear significant. These results are in agreement with earlier work (Cook, 1967; Ptitsyn, 1969; Chou & Fasman, 1973).

The inner part of the helix contains two kinds of segments of a very distinct composition and they are long enough to be considered as different entities. On this basis, we can define: (i) *hydrophobic clustered-zones*. These correspond to internal segments (q residues as an average) in which there is one, or several, hydrophobic triplets of the type 1-2-5 or 1-4-5 and in addition, adjacent groups of residues mainly hydrophobic ($\geq$50%). (ii) *Hydrophobic-depleted zones*. These are internal segments (6 residues on average) that are relatively poor in hydrophobic residues. (iii) *N-terminal triplets*. (iv) *C-terminal triplets*. In some cases, small overlapping of segments classified as (i) or (ii) with the terminal triplets exists.

Table 1 details the segments for each protein, considered as clustered and depleted zones. There are 45, 17, 63 and 63 segments (or 391, 105, 189 and 189 residues) which correspond, respectively, to zones classified as (i), (ii), (iii) and (iv). Table 2 shows the amino acid composition of the four zones and also an analysis of the reliability of deviations in amino acid composition of the four α-helical zones as compared with the whole α-helical region. The χ^2 test was applied to compare the whole α-helical region with its zones. The hydrophobic-clustered zone is different from the α-helix with a 99·9% reliability and the N-terminal and C-terminal triplets with a 99% reliability. The hydrophobic-depleted zones do not show a general difference from the whole α-helical region.

From the data reported in Table 2 we can draw the following conclusions: (1) Ile,

TABLE 1

Hydrophobic-clustered and hydrophobic-depleted zones found in the 12 proteins as defined by the criteria of enchaining of hydrophobic triplets

Protein	Hydrophobic-clustered			Hydrophobic-depleted		
Ferricytochrome C[a]		94–101			—	
Ribonuclease A[b]		—		7–10	30–33	53–56
Lysozyme[c, d]	8–12		25–32		90–94	
Myoglobin[e, f]	7–17 86–90	28–33 103–115	66–76 127–138 142–146	22–27		60–65
Subtilisin BPN'[g, h]	68–73 133–144	104–108 227–235	111–116 269–273		244–250	
α-Haemoglobin[e]	10–14 82–86	24–33 106–111	62–69 123–136	5–9		96–105
α-Chymotripsin[i, j]		229–233		162–167		234–238
Carboxypeptidase A[k, l]	15–17 99–103	73–77 176–183 286–290	82–86 219–230 294–301		114–120	256–260
Papain[m, n]	26–38	71–78	120–126		51–55	
Staphylococcal nuclease[o]	99–104				56–65	124–132
Thermolysin[p]	70–87 236–244 292–296	137–151 263–274 306–313	163–180 284–289		—	
Myogen[q]	9–15	29–33 102–106	43–50		80–36	

References : [a]Dickerson *et al.* (1971); [b]Kartha *et al.* (1967); [c]Blake *et al.* (1965); [d]Ramachandran & Sasisekharan (1968); [e]Dickerson & Geis (1969); [f]Kendrew *et al.* (1961); [g]Wright *et al.* (1969); [h]Kraut (1971); [i]Matthews *et al.* (1967); [j]Blow (1971); [k]Bradshaw *et al.* (1969); [l]Hartsuck & Lipscomb (1971); [m]Drenth *et al.* (1969); [n]Drenth *et al.* (1971); [o]Cotton & Hazen (1971); [p]Colman *et al.* (1972); [q]Nockolds *et al.* (1972).

TABLE 2

Distribution of amino acid residues in different helical zones and the significance of their deviations as compared with the total helical region

	α-Helical region		Hydrophobic-clustered			Hydrophobic-depleted			N-terminal triplets			C-terminal triplets		
	No.	%	No.	%	signif.	No.	%	signif.	No.	%	signif.	No.	%	signif.
Leu	71	8·98	55	14·07	96·41+	9	8·57	52·79	9	4·76	97·98−−	15	7·94	63·31
Ile	39	4·93	35	8·95	99·99+++	0	0·00	98·93−−	10	5·29	58·32	5	2·64	90·66
Val	61	7·72	45	11·51	99·79+++	4	3·81	92·35	10	5·29	89·62	16	8·47	70·54
Met	10	1·26	6	1·53	69·15	2	1·90	72·90	2	1·06	60·26	2	1·06	57·93
Phe	30	3·79	18	4·60	80·78	6	5·71	85·77	4	2·12	88·88	7	3·70	51·60
Tyr	28	3·54	20	5·12	95·73+	2	1·90	79·95	6	3·17	61·03	9	4·76	84·38
Trp	13	1·64	10	2·56	92·36	1	0·95	70·54	3	1·59	54·38	1	0·53	92·07
Ala	119	15·06	83	21·23	99·97+++	11	10·48	89·07	24	12·70	82·38	27	14·29	52·79
Thr	44	5·56	16	4·09	89·25	8	7·62	83·65	11	5·82	55·57	8	4·23	74·86
Cys	9	1·13	1	0·26	94·41	4	3·81	99·52+++	3	1·59	71·90	3	1·59	73·89
Ser	52	6·58	12	3·07	99·70−−−	8	7·62	68·08	16	8·47	84·85	13	6·88	61·79
Gln	37	4·68	8	2·05	99·22−−	6	5·71	97·38++	8	4·23	61·79	14	7·41	97·06+
Asn	32	4·05	9	2·30	95·73−	6	5·71	81·33	12	6·35	94·41	10	5·29	83·40
His	31	3·92	5	1·28	99·61−−−	5	4·76	69·15	7	3·70	56·36	11	5·82	92·65
Lys	62	7·84	21	5·37	96·16−	12	11·43	92·22	7	3·70	98·34−−	22	11·64	98·17++
Arg	16	2·02	3	0·77	96·33−	4	3·81	90·66	3	1·59	66·64	5	2·64	75·17
Glu	43	5·44	13	3·32	96·49−	6	5·71	56·75	12	6·35	70·54	6	3·17	89·44
Asp	38	4·81	17	4·35	64·80	5	4·76	51·99	16	8·46	99·05++	7	3·70	72·24
Gly	39	4·93	12	3·07	95·15−	5	4·76	51·60	14	7·41	94·06	8	4·23	64·43
Pro	16	2·02	2	0·51	98·17−−	1	0·95	77·64	13	6·88	≥99·99+++	0	0·00	97·19−−

Significance of deviation, positive or negative: ±±±, greater than 99·5; ±±, greater than 97·5%; ±, greater than 95%.

Ala and Val, and to a lesser extent Leu and Tyr tend to accumulate in hydrophobic-clustered zones. (2) Ala and Leu, reported as significantly abundant in α-helical regions, are also spread among the other three zones, although it should be noted that N-terminal zones have a Leu-depleted character. (3) His and Glu, the other two residues which predominate in α-helices, are preferentially spread in zones other than hydrophobic-clustered ones. (4) Ser, Asn and Gly are rarer in hydrophobic-clustered zones than in the whole α-helical region. (5) Cys and Gln are preferentially located in hydrophobic-depleted zones. (6) It is interesting to note the complete absence of Ile within hydrophobic-depleted zones. (7) There is a greater abundance of Lys and Gln in the C-terminal zones. (8) The N-terminal zone of an α-helix tends to accumulate Asp and Pro, and to be depleted of Leu and Lys.

(b) *Medium-range interactions of hydrophobic arcs*

Perutz *et al.* (1965) observed the presence of a regular periodicity of sites occupied by non-polar residues along each of the longer α-helical regions of haemoglobin. This remark correlates with the suggestion that intrahelical interactions involve residues n and $n \pm 3$ or $n \pm 4$ (Low & Edsall, 1956), such interactions being helix-stabilizing when the residues concerned are hydrophobic (Némethy & Scheraga, 1962). In connection with the work of Perutz *et al.* (1965), Cook (1967) refers to a personal communication from Guzzo, who suggested that hydrophobic residues in a 1-2-5 or 1-4-5 relation are necessary for helix formation.

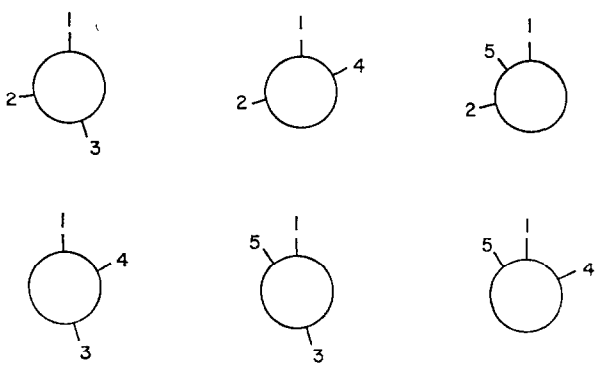

FIG. 1. Schematic transverse section of α-helices showing the angular distribution of residues in each of the 6 triplets found within a helical pentapeptide.

In the second part of our paper, emphasis is placed on such ideas and we present a statistical study of the occurrence of hydrophobic triplets of six different types, corresponding to those which can be considered in a pentapeptide (1-2-3, 1-2-4, 1-2-5, 1-3-4, 1-3-5 and 1-4-5). In Figure 1 six helical wheels are shown with the relative orientations of the different triplets. It can be seen that the angle between the two more distant residues in each wheel is about 100° for the triplets 1-2-5 and 1-4-5, whereas for the other four triplets it is about 180°. In consequence, the relative closeness of the triplet residues is highest for positions 1-2-5 and 1-4-5, and medium-range interactions can be facilitated by this proximity.

A comparison between the theoretical probability and the experimental frequency for hydrophobic triplets is given in Figure 2. In the upper part of the Figure it is shown that deviations in the α-helical regions are more than $+2\sigma$ for triplets 1-4-5 and 1-2-5, and less than $\pm\sigma$ for the other four triplets. A similar consideration can be made for non-helical regions. Although deviations are less marked in this case, in the lower part of Figure 2 negative deviations greater than σ can be seen for triplets 1-4-5 and 1-2-5, and deviations less than σ for the other triplets. From these results there is a clear preference for triplets 1-4-5 and 1-2-5 to be concentrated in α-helical regions. It is also interesting to note, by observing the sequence of the proteins studied, that quite a number of these triplets are enchained, giving rise to relatively large regions defined by us as hydrophobic-clustered regions. A number of 1-2-5 and 1-4-5 hydrophobic triplets are found in non-helical regions. Their presence can be explained by several facts. They are generally found isolated (i.e. not enchained to other hydrophobic triplets) and are quite often surrounded by helix-breaking residues. Furthermore, the limitation of the size of our sample does not allow classification of the 512 ($8\times8\times8$) different hydrophobic triplets in terms of their helix-stabilizing ability. It is possible that some of these triplets can be defined as helix-indifferent. On the other hand, the influence of long-range interactions in disrupting helices must be taken into account.

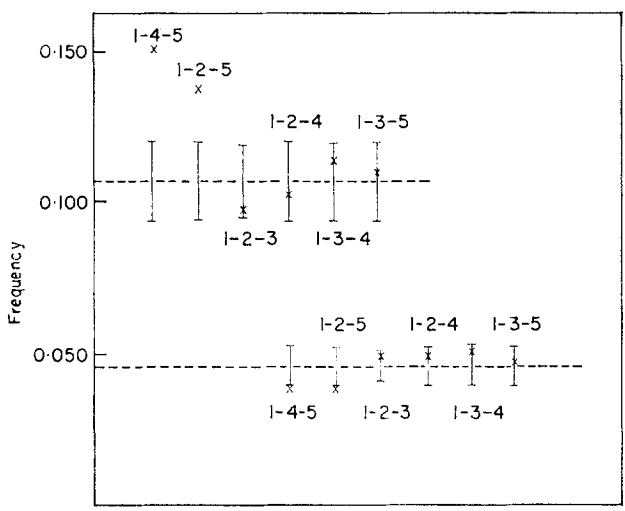

FIG. 2. Hydrophobic triplet frequencies in α-helical regions (upper part of the Figure), and in non-helical regions (lower part of the Figure). Vertical bars represent the calculated standard deviations, and broken horizontal lines show the probability (in frequency values) of finding hydrophobic triplets, for the two regions considered.

The dependence of the results of the triplet analyses on the size of the protein sample was tested by two different methods as described in Materials and Methods, section (d).

In the first of these tests, the variation of the normalized standard deviation for each hydrophobic triplet as a function of the order of combination of the 12 proteins is shown in Figure 3 for α-helical regions and in Figure 4 for non-helical regions. If the lower combinatory orders are not considered, the standard deviation decreases

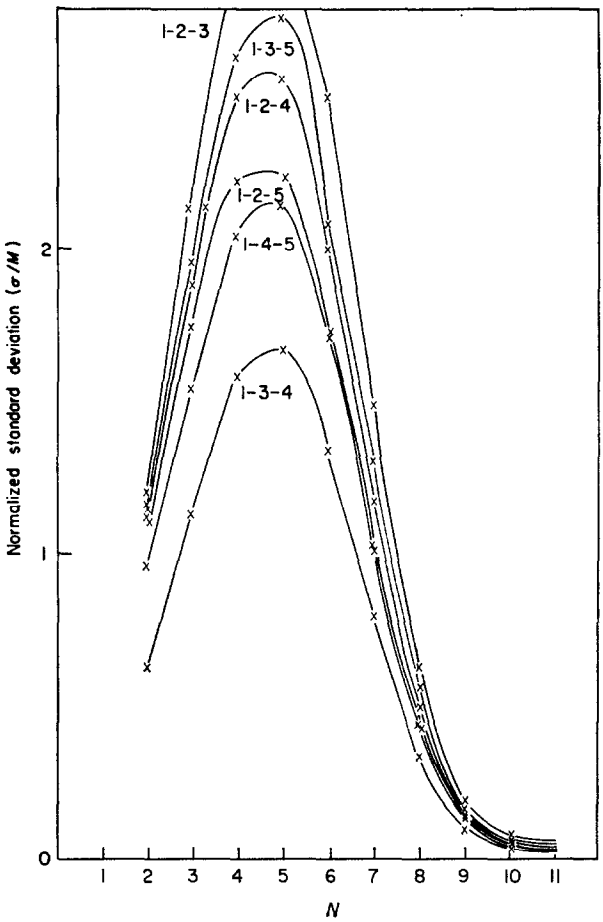

FIG. 3. Normalized standard deviation for the different hydrophobic triplets in α-helical regions as a function of the order (N) of combination of proteins. Normalization was performed with respect to the average value of frequencies of all combinations found in each combinatory order.

as the order of combinations increases, and above an order of nine (or eight in some cases) the values of σ/M remain low. The test shows that if our analyses were carried out with any combination of eight or nine of the 12 proteins chosen, the results would be significantly the same.

The second test enables us to confirm that any combination of eight proteins would give the same result. The correlation coefficients obtained are consistently high. Only 11% of scattered values are below 0·9.

4. Discussion

Kotelchuck & Scheraga (1968) examined the importance of the interaction of the side chain with the amide group of the same peptide unit (—CHR—CO—NH—). This type of short-range interaction would be primarily responsible for the α-helix formation if the single peptide units considered presented minimum energy conformations at ϕ and ψ values characteristic of α-helices (Kotelchuck & Scheraga, 1969).

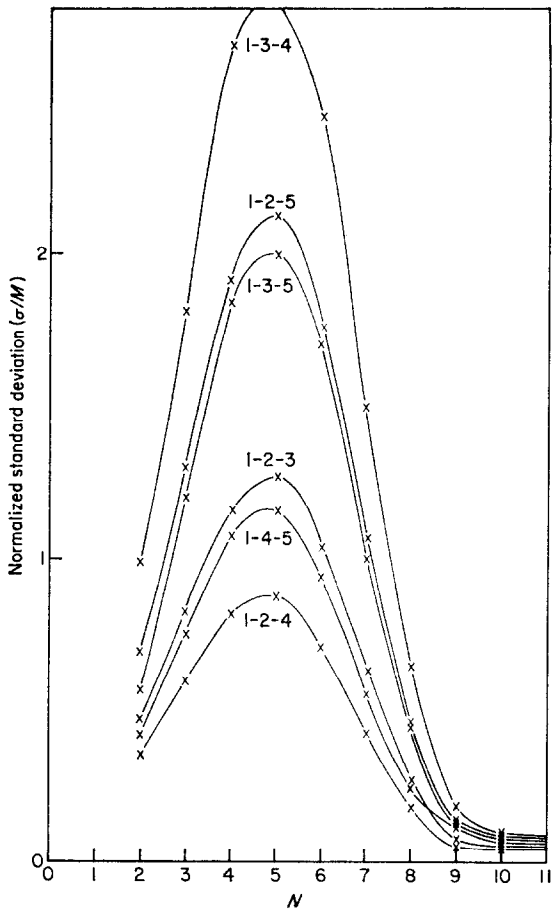

FIG. 4. Normalized standard deviation for the different hydrophobic triplets in non-helical regions as a function of the order (N) of combination of proteins. Normalization was performed with respect to the average value of frequencies of all combinations found in each combinatory order.

It has been reported (Kotelchuck & Scheraga, 1969; Robson & Pain, 1971; Leberman, 1971) that helix initiation at the N-terminal site is directed by helix-forming residues taken singly. Our studies on amino acid distribution, and those carried out by other authors, show that even in the central core, there is accumulation of helix-forming residues, this effect being increased at the hydrophobic-clustered zones, and less defined at the hydrophobic-depleted zones. From these considerations, it can be established that helix *nucleation* may be codified by the residues taken independently of each other. We define the term nucleation as the tendency of the peptide units involved to adopt characteristic ϕ and ψ values. Helix nucleation would be, therefore, the effect of short-range interactions. Discrepancies between helices found in proteins and those predicted by the empirical rules of best-fitting may be explained by considering the influence of medium and long-range interactions.

The effect of medium-range interactions (i.e. interactions among residues n, $n \pm 1$, $n \pm 2$, $n \pm 3$ and $n \pm 4$) on regions of helical potential can be studied by considering doublets, triplets or even multiplets. Extensive work has been carried

out on doublet distribution analyses (Periti *et al.*, 1967; Kotelchuck *et al.*, 1969; Robson & Pain, 1971; Finkelstein & Ptitsyn, 1971; Nagano, 1973) and the main conclusions that have been drawn are: (1) doublets by themselves do not seem to have any fundamental role in forming the helices; (2) there is a significant concentration of helix-breaking doublets at both sides of the C-terminal of the helix. Therefore it can be established that medium-range interactions derived from doublets are mainly helix-destabilizing.

The triplet distribution analyses of hydrophobic residues, described in this paper, can be considered as a statistical demonstration of the observations on multiplet distribution reported by Perutz *et al.* (1965), and Schiffer & Edmunson (1967). From our results it can be concluded that interacting hydrophobic triplets (1-2-5 and 1-4-5) are helix-stabilizing.

Wu & Kabat (1973) reported an interesting study on the effect of nearest-neighbour residues, $n \pm 1$, on ϕ and ψ values of residues n. This can be considered as an effect of 1-2-3 triplets on helicity. However, the number of possible triplets (8000) compared with the number within the proteins studied (1561) indicates that limitations exist for verifying the results. Triplet distribution analysis can also be extended to different groupings of residues (48,000 for the six types of triplets found in a pentapeptide) but the small sample of proteins available at present do not allow serious statistical studies. Therefore, in order to find new types of triplet interactions, an appropriate strategy of residue grouping will have to be established.

It seems convenient to make a comment on the effects resulting from long-range interactions. Such interactions can take place among residues located at different parts of a protein; they can also originate from internal tensions produced by the folding of the molecule, from Cys–Cys bridging, from the proximity of β-structures and reverse turns to the α-helix, and from the prosthetic groups when they are present. The pooling of these effects can co-operate and give rise to "mistakes" when the structural translation takes place. The evaluation of long-range interactions represents an extremely complex problem and, therefore, the proposal of Robson & Pain (1971) of considering their effects as informational noise should be accepted as a limitation due to the high conformational complexity of proteins (Phillips, 1966). An estimate of such noise at particular sites of the molecule is far from being established.

In conclusion, we would like to draw a picture of the influence of the different forces acting on the secondary structure of proteins. The process can be considered as a sophisticated translational pathway, which starts from primary structure and goes to secondary structure, then continues to structures of higher order and finally returns to the secondary structure. The code that governs the translation may be described as a three-step informational flow. A first set of orders, derived from short-range interactions, gives rise to the nucleation of helices or other structures in a definite zone of a protein. A second set of orders, derived from medium-range interactions among groups of residues—such as disrupting doublets or stabilizing hydrophobic triplets—defines the helical regions. Finally, a third set of informational orders, derived from long-range interactions, modulates the secondary structure within the folded protein.

The authors cordially thank Drs L. Puigjaner and R. Carbó for consultations on the statistical aspects of this problem. The authors are also indebted to Dr Ellis for revising the English.

REFERENCES

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature (London)*, **206**, 757–761.

Blow, D. M. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 185–212, 3rd edn, Academic Press, New York.

Bradshaw, R. A., Ericsson, L. H., Walsh, K. A. & Neurath, H. (1969). *Proc. Nat. Acad. Sci., U.S.A.* **63**, 1389–1394.

Chou, P. Y. & Fasman, G. D. (1973). *J. Mol. Biol.* **74**, 263–281.

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70**, 701–724.

Cook, D. A. (1967). *J. Mol. Biol.* **29**, 167–171.

Cotton, F. A. & Hazen, E. S. Jr. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 4, pp. 153–175, 3rd edn, Academic Press, New York.

Crawford, J. L., Lipscomb, W. N. & Schellman, C. G. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 538–542.

Dickerson, R. E. & Geis, I. (1969). In *The Structure and Action of Proteins*, p. 52, Harper & Row, New York.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., Cooper, A. & Margoliash, E. (1971). *J. Biol. Chem.* **246**, 1511–1535.

Drenth, J., Jansonius, J. N., Koekoek, R., Swen, H. M. & Wolthers, B. G. (1969). *Nature (London)*, **218**, 929–932.

Drenth, J., Jansonius, J. N., Koekoek, R. & Wolthers, B. G. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 485–499, 3rd edn, Academic Press, New York.

Finkelstein, A. V. & Ptitsyn, O. B. (1971). *J. Mol. Biol.* **62**, 613–624.

Guzzo, A. V. (1965). *Biophys. J.* **5**, 809–822.

Hartsuck, J. A. & Lipscomb, W. N. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 1–56, Academic Press, New York.

Kartha, G., Bello, J. & Harker, D. (1967). *Nature (London)*, **213**, 862–865.

Kendrew, J. C., Watson, H. C., Strandberg, B. E., Dickerson, R. E., Phillips, D. C. & Shore, V. C. (1961). *Nature (London)*, **190**, 666–670.

Kotelchuck, D. & Scheraga, H. A. (1968). *Proc. Nat. Acad. Sci., U.S.A.* **61**, 1163–1170.

Kotelchuck, D. & Scheraga, H. A. (1969). *Proc. Nat. Acad. Sci., U.S.A.* **62**, 14–21.

Kotelchuck, D., Dygert, M. & Scheraga, H. A. (1969). *Proc. Nat. Acad. Sci., U.S.A.* **63**, 615–622.

Kraut, J. (1971). In *The Enzymes* (Boyer, P. D., ed.), vol. 3, pp. 547–560, 3rd edn, Academic Press, New York.

Krigbaum, W. R. & Rubin, B. H. (1971). *Biochim. Biophys. Acta*, **229**, 368–383.

Leberman, R. (1971). *J. Mol. Biol.* **55**, 23–30.

Low, B. W. & Edsall, J. T. (1956). In *Currents in Biochemical Research 1956* (Green, D. E., ed.), pp. 378–433, Interscience Publishers, New York.

Low, B. W., Lovell, F. M. & Rudko, A. D. (1968). *Proc. Nat. Acad. Sci., U.S.A.* **60**, 1519–1526.

Matthews, B. W., Sigler, P. B., Henderson, R. & Blow, D. M. (1967). *Nature (London)*, **214**, 652–656.

Nagano, K. (1973). *J. Mol. Biol.* **75**, 401–420.

Némethy, G. & Scheraga, H. A. (1962). *J. Phys. Chem.* **66**, 1773–1789.

Nockolds, C. E., Kretsinger, R. H., Coffee, C. J. & Bradshaw, R. A. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **69**, 581–584.

Pain, R. H. & Robson, B. (1970). *Nature (London)*, **227**, 62–63.

Periti, P. F., Quagliarotti, G. & Liquori, A. M. (1967). *J. Mol. Biol.* **24**, 313–322.

Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). *J. Mol. Biol.* **13**, 669–678.

Phillips, D. C. (1966). *Sci. Amer.* pp. 78–90.

Ponnuswammy, P. K., Warme, P. K. & Scheraga, H. A. (1973). *Proc. Nat. Acad. Sci., U.S.A.* **70**, 830–833.

Prothero, J. W. (1966). *Biophys. J.* **6**, 367–370.
Ptitsyn, O. B. (1969). *J. Mol. Biol.* **42**, 501–510.
Ramachandran, G. N. & Sasisekharan, V. (1968). *Advan. Protein Chem.* **23**, 283–437.
Robson, B. & Pain, R. H. (1971). *J. Mol. Biol.* **58**, 237–259.
Schiffer, M. & Edmunson, A. B. (1967). *Biophys. J.* **7**, 121–135.
Wright, C. S., Alden, P. A. & Kraut, J. (1969). *Nature (London)*, **221**, 235–242.
Wu, T. T. & Kabat, E. A. (1971). *Proc. Nat. Acad. Sci., U.S.A.* **68**, 1501–1506.
Wu, T. T. & Kabat, E. A. (1973). *J. Mol. Biol.* **75**, 13–31.