A MODEL OF ALPHA-HELICAL DISTRIBUTION IN PROTEINS

JOHN W. PROTHERO

From the Department of Biological Structure, University of Washington, Seattle, Washington 98105

ABSTRACT It is shown that alpha-helical content of eleven proteins is well correlated with alanine plus leucine content. These residues, taken singly or together, are to a first approximation randomly distributed in the four proteins whose tertiary structures have been determined (i.e., myoglobin, lysozyme, ribonuclease, α-chymotrypsin). A model based on the concept that certain randomly distributed residues specifically participate in helix nucleation is shown to be in reasonable agreement with the presently published structures.

INTRODUCTION

The general problem of predicting the secondary structure of a protein from the amino acid sequence continues to be of interest. Attention is focused particularly on α-helical secondary structure as this appears to occur in a wide variety of proteins. Experimental data on the occurrence of α-helical sequences in globular proteins are derived from combined amino acid sequence studies and single crystal X-ray studies. Such data are now available for myoglobin (1, 2), lysozyme (3, 4), ribonuclease (5, 6), α-chymotrypsin (7, 8), and, in part by inference, for α-, β-, and γ-hemoglobin (9–11). Useful data are also accumulating from optical and X-ray studies of synthetic polypeptides in solution (12–15). Two general approaches to the analysis of α-helical distribution may be taken. In one, which may be termed the "fundamental approach," the analysis proceeds from considerations of physical chemistry and statistical mechanics (16–22). In the second, which may be termed the "empirical approach," the analysis proceeds directly from a study of the experimental data (23–32). The present paper takes the second approach.

In carrying out an empirical analysis of α-helical distribution, it is useful to keep the present theoretical picture, as derived from fundamental studies, in mind. In general, the secondary structure of a protein is known to be strongly dependent on the solvent present. The transition from a coiled structure to an α-helical structure, say, is considered to be a particular example of a cooperative process. The study of cooperative processes usually begins with the Ising model (33). Ising approached

the general problem of phase transitions from the statistical mechanical viewpoint. The fundamental problem is to calculate correlation functions between nearest neighbors and infinitely separated neighbors.

In applying the Ising approach to the problem of the helix-coil transition, it has proved useful to think of the "nucleation" of helical segments and their subsequent "growth" (16–22). The general notion of nucleation and growth is employed explicitly in the model considered below.

ANALYSIS OF DATA

Before describing the model, it is advantageous to examine the data for those proteins whose primary, secondary, and tertiary structure is now known. Table I contains a summary of the number of times each residue occurs in a helical or nonhelical region for each of seven proteins. There are over 1000 residues, which are almost evenly divided between helical and nonhelical regions. It will be seen that some residues, such as alanine (Ala) occur twice as often in helical as in nonhelical regions, whereas the reverse is true for other residues such as threonine (Thr).

It is also useful to examine the data derived simply from amino acid compositions and optical rotatory dispersion measurements, as was first done by Davies (23). That is, for a given protein, the α-helical content is known, the amino acid composition is known, but the primary structure (in general) is not known. The data selected from that compiled by Davies refers to those proteins which are presumably not homologous. These data have been plotted for each individual amino acid as a function of the helical content of the respective proteins in Figs. 1, 2, and 3. Each point on each plot represents the amount of a given residue in a protein having a certain helical content (see legend, Fig. 1 for list of proteins). The residues denoted by Asn, Asp, Gln, Glu, CyS have been plotted for four proteins only, since they have not been determined separately for the other proteins. Two plots are shown for combined residues, namely, alanine plus leucine, and serine plus threonine (cf. Fig. 3). Note that equal numbers of residues are positively and negatively correlated with helical content (i.e., 10 each).

These data (i.e. Table I and Figs. 1–3) may be utilized in a variety of ways. Mainly, we are interested in ranking the various residues according to their tendency to occur in helical regions. Since the data are primarily derived from seven proteins (Table I), it is useful to ask how typical of proteins in general this sample may be. An estimate may be obtained by ranking the residues according to their frequency of occurrence and comparing the resultant order with that calculated from a larger (random?) sample (34). Such a comparison is given in the top two rows of Table II. It will be seen that while the observed sample is generally similar to the random population, the sample appears to be biased in favour of valine and histidine and deficient in arginine and methionine.

In Table II the residues have also been ranked according to four different criteria.

TABLE I

OCCURRENCE OF RESIDUES IN HELICAL REGIONS

Protein	Res	Ala	Arg	Asn	Asp	CySH	CyS	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Mb	Hel	15	3	1	5	0	0	4	12	7	7	8	15	13	2	3	3	6	4	2	2	7	119
	Nhel	2	1	1	1	0	0	1	2	4	5	1	3	6	0	3	1	0	1	0	1	1	34
Ly	Hel	10	5	3	1	0	5	1	2	2	1	2	4	2	1	1	0	6	1	4	0	5	56
	Nhel	2	6	10	7	0	3	2	0	10	0	4	4	4	1	2	2	4	6	2	3	1	73
RNAse	Hel	4	2	1	1	0	1	3	1	0	1	0	2	2	2	1	0	1	0	0	0	2	24
	Nhel	8	2	8	5	0	7	4	4	3	3	3	0	8	2	2	4	14	10	0	6	7	100
Chym	Hel	2	0	1	0	0	0	2	0	0	0	0	1	0	0	0	0	0	1	0	0	1	8
	Nhel	20	3	13	8	0	10	8	4	28	2	10	18	14	2	6	9	27	21	8	4	22	237
Hb-α	Hel	16	2	0	7	1	0	0	4	7	7	0	13	10	1	5	3	9	8	1	3	11	108
	Nhel	5	1	0	5	0	0	0	1	0	3	0	5	1	1	2	4	2	1	0	0	2	33
Hb-β	Hel	15	3	0	9	2	0	0	9	10	7	0	14	8	1	4	5	3	3	2	3	15	113
	Nhel	0	0	0	4	0	0	0	2	3	2	0	5	2	0	4	2	2	4	0	0	3	33
Hb-γ	Hel	11	3	0	8	1	0	0	10	9	5	4	14	9	2	3	3	8	7	3	2	11	113
	Nhel	0	0	0	5	0	0	0	2	4	2	0	3	3	0	5	1	3	3	0	0	2	33
Subtotals	Hel	73	18	6	31	4	6	10	38	35	28	14	63	44	9	17	14	33	24	12	10	52	541
	Nhel	37	13	32	35	0	20	15	15	52	17	18	38	38	6	24	23	52	46	10	14	38	543
Totals		110	31	38	66	4	26	25	53	87	45	32	101	82	15	41	37	85	70	22	24	90	1084

A summary of the number of times each residue occurs in helical (Hel) and nonhelical (Nhel) regions for the proteins myoglobin (Mb), lysozyme (Ly), ribonuclease (RNAse), α-chymotrypsin (Chym), and the α, β, and γ hemoglobins (Hb-α, Hb-β, Hb-γ).

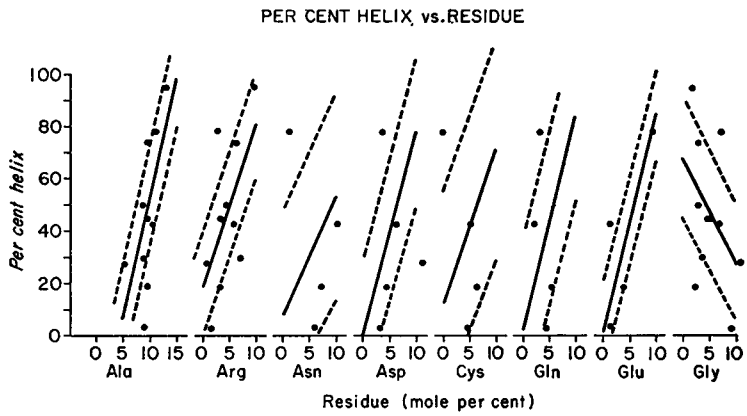

FIGURE 1 A plot of helical content in various proteins vs. the mole per cent of individual residues. The proteins include either chymotrypsinogen A, ribonuclease, pepsin, tobacco mosaic virus, lysozyme, ovalbumin, heavy meromyosin, bovine serum albumin, myoglobulin and paramyosin, or in some cases only myoglobin, lysozyme, ribonuclease, and chymotrypsinogen A. Data adapted from that given by Davies (18). The solid line represents the least squares regression equation and the broken line represents one standard deviation.

In general, the residue r_i precedes the residue r_j if the coefficient c_i is greater than the coefficient c_j, where the coefficients c_i, c_j are calculated according to one of the four following relations.

For the residue r_i the coefficient c_i is given by:

Rank 1 $c_i = [(\text{Hel})/(1084)] \times 100$

Rank 2 $c_i = [(\text{Hel})/(\text{NHel})] \times 100$

Rank 3 $c_i = (\text{Hel}/\text{NHel})(\text{Hel} + \text{NHel})$ if $\text{Hel} > \text{NHel}$

$c_i = -(\text{NHel}/\text{Hel})(\text{Hel} + \text{NHel})$ if $\text{Hel} < \text{NHel}$

Rank 4 $c_i =$ correlation coefficient of regression lines shown in Figs. 1–3.

where Hel (NHel) is the number of times a residue occurs in a helical (nonhelical) region (1084 is the total number of residues appearing in Table I).

The first method, rank 1, is simply a measure of how often a given residue occurs in helical regions. Thus, alanine occurs more frequently than any other residue in helical regions (see Table I).

The second method, rank 2, is a measure of how often a given residue occurs in helical regions as compared to nonhelical regions, without regard to how often that residue occurs in the whole population. This method of ranking can give a possibly spurious result, as in the case of CySH, which occurs four times in helical regions and zero times in nonhelical regions, thus giving an infinite value of c_i. The third method is basically similar to the second, but has the effect of forcing those resi-

dues which are negatively correlated with helical regions over to the far right. The fouth method is based on the data adapted from that assembled by Davies and plotted in Figs. 1–3. The residues have been ranked according to the correlation coefficient associated with the regression line for each plot.

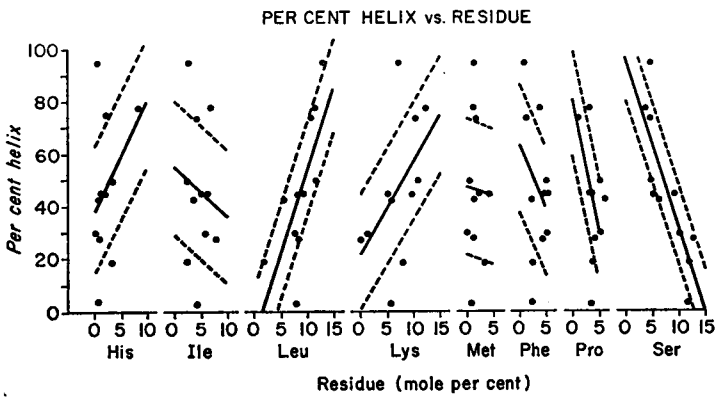

FIGURE 2 See legend for Fig. 1.

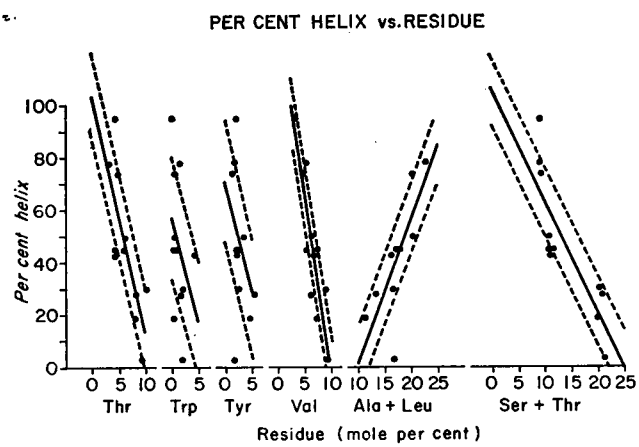

FIGURE 3 See legend for Fig. 1.

It is of interest that regardless of the method by which the residues are ranked, certain residues consistently occur at the far left of Table II (i.e., are highly correlated with α-helical distribution). Thus alanine, leucine, and glutamate occur in the first five residues in each case. The presence of CySH at the far left is doubtless spurious, as indicated above. A similar statement cannot be made about any residue on the right, although Asn, Thr, Ser, and CyS occur there frequently. Valine is anomalous, in that it occurs on the left twice, roughly in the middle once (rank 2), and on the right once (rank 4). It is doubtful if any significance can be attached to the precise ordering of the residues. The pattern, however, is likely to be significant.

TABLE II

RESIDUES ORDERED WITH RESPECT TO HELICAL DISPOSITION

Random	Ala	Glx	Asx	Leu	Gly	Lys	Ser	Val	Arg	Thr	Pro	Ile	Met	Phe	Tyr	Cyx	Trp	His			
Observed	Ala	Asx	Leu	Val	Gly	Ser	Lys	Glx	Thr	His	Phe	Pro	Ile	Arg	Cyx	Tyr	Trp	Met			
Rank 1	Ala	Leu	Val	Lys	Glu	Gly	Ser	Asp	His	Thr	Arg	Phe	Ile	Pro	Trp	Tyr	Gln	Met	Asn	CyS	CySH
Rank 2	CySH	Glu	Ala	Leu	His	Met	Arg	Val	Trp	Lys	Asp	Ile	Tyr	Phe	Gly	Gln	Ser	Pro	Thr	CyS	Asn
Rank 3	CySH	Ala	Leu	Glu	Val	Lys	His	Arg	Met	Trp	Tyr	Ile	Gln	Phe	Pro	Asp	CyS	Gly	Thr	Ser	Asn
Rank 4	Leu	Glu	Ala	Arg	Lys	His	Asp	Gln	Asn	CyS	Met	Ile	Phe	Trp	Tyr	Gly	Pro	Thr	Ser	Val	

The top row gives the frequency of occurrence of residues in a random sample of proteins (31). The second row gives the observed frequency in the seven proteins analyzed in Table I. The bottom four rows rank the various residues according to their tendency to occur in helical regions, the highest tendency being to the left. For method of ranking see text.

TABLE III

DISTRIBUTION OF ALANINE IN FOUR PROTEINS

	Mb				Ly				RNAse				Chym			
Res/span Span	1		2		1		2		1		2		1		2	
	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc
2	15	14 ± 0.4	1	1.5 ± 0.1	8	10 ± 0.4	2	0.9 ± 0.1	8	10 ± 0.4	2	1.8 ± 0.2	14	18.4 ± 0.4	4	1.7 ± 0.1
3	15	12 ± 0.5	1	2.0 ± 0.2	8	9 ± 0.5	2	1.3 ± 0.2	7	9 ± 0.5	1	1.0 ± 0.2	17	16.9 ± 0.5	2	2.3 ± 0.2
4	13	11 ± 0.5	2	2.4 ± 0.3	7	8 ± 0.5	1	1.5 ± 0.2	8	8 ± 0.5	2	1.3 ± 0.2	14	15.4 ± 0.5	4	2.8 ± 0.2
5	11	10 ± 0.6	3	2.7 ± 0.3	6	8 ± 0.6	3	1.8 ± 0.3	7	7 ± 0.5	2	1.6 ± 0.3	14	14.1 ± 0.5	4	3.2 ± 0.3
6	11	9 ± 0.6	3	2.9 ± 0.3	7	7 ± 0.6	1	1.9 ± 0.3	6	7 ± 0.6	1	1.8 ± 0.3	10	12.9 ± 0.6	5	3.5 ± 0.3
7	9	8 ± 0.6	4	3.1 ± 0.4	5	6 ± 0.6	2	2.0 ± 0.3	4	6 ± 0.6	2	2.0 ± 0.3	10	11.8 ± 0.6	6	3.7 ± 0.3
8	7	7 ± 0.6	5	3.2 ± 0.4	3	6 ± 0.6	3	2.1 ± 0.4	4	6 ± 0.6	2	2.1 ± 0.3	12	10.8 ± 0.6	4	3.9 ± 0.4
9	8	6 ± 0.6	3	3.2 ± 0.4	6	5 ± 0.6	3	2.2 ± 0.4	6	5 ± 0.6	1	2.2 ± 0.4	9	9.9 ± 0.6	6	4.0 ± 0.4
10	4	6 ± 0.6	5	3.2 ± 0.5	3	5 ± 0.6	4	2.2 ± 0.4	2	5 ± 0.6	3	2.2 ± 0.4	11	9.1 ± 0.6	3	4.1 ± 0.4

The observed distribution of alanine in the four proteins myoglobin (Mb), lysozyme (Ly), ribonuclease (RNAse) and α-chymotrypsin (Chym) obtained by dividing each protein into successive segments of length 2–10 (see Span) and counting the number of times one and only one or two and only two residues occur in a single segment. The calculated columns were derived from a Poisson distribution (see text).

TABLE IV

DISTRIBUTION OF LEUCINE IN FOUR PROTEINS

Res/span Span	Mb				Ly				RNAse				Chym			
	1		2		1		2		1		2		1		2	
	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc
2	18	14.3 ± 0.4	0	1.7 ± 0.1	5	7.2 ± 0.3	1	0.4 ± 0.1	2	1.9 ± 0.2	0	0.0	17	16.3 ± 0.4	1	1.3 ± 0.1
3	18	12.7 ± 0.5	0	2.2 ± 0.2	6	6.9 ± 0.4	1	0.5 ± 0.1	2	1.9 ± 0.2	0	0.0	15	15.1 ± 0.4	2	1.8 ± 0.1
4	12	11.3 ± 0.5	3	2.7 ± 0.3	5	6.5 ± 0.4	1	0.7 ± 0.1	2	1.9 ± 0.2	0	0.1 ± 0.1	14	14.0 ± 0.5	1	2.2 ± 0.2
5	16	10.1 ± 0.6	1	3.0 ± 0.3	5	6.2 ± 0.5	1	0.8 ± 0.2	2	1.8 ± 0.3	0	0.1 ± 0.1	15	13.0 ± 0.5	2	2.5 ± 0.2
6	10	8.9 ± 0.6	4	3.2 ± 0.4	5	5.9 ± 0.5	1	0.9 ± 0.2	2	1.8 ± 0.3	0	0.1 ± 0.1	13	12.0 ± 0.5	1	2.8 ± 0.3
7	7	8.0 ± 0.6	5	3.3 ± 0.4	5	5.6 ± 0.6	1	1.0 ± 0.2	2	1.8 ± 0.3	0	0.1 ± 0.1	11	11.1 ± 0.6	4	3.0 ± 0.3
8	12	7.1 ± 0.6	3	3.3 ± 0.4	5	5.3 ± 0.6	1	1.1 ± 0.3	2	1.8 ± 0.3	0	0.1 ± 0.1	9	10.3 ± 0.6	3	3.2 ± 0.3
9	10	6.3 ± 0.6	4	3.4 ± 0.4	5	5.0 ± 0.6	1	1.2 ± 0.3	2	1.7 ± 0.4	0	0.1 ± 0.1	11	9.6 ± 0.6	1	3.3 ± 0.3
10	6	5.7 ± 0.6	6	3.3 ± 0.5	5	4.8 ± 0.6	1	1.2 ± 0.3	2	1.7 ± 0.4	0	0.1 ± 0.1	11	8.9 ± 0.6	2	3.4 ± 0.4

See legend, Table III.

TABLE V

DISTRIBUTION OF ALANINE-LEUCINE IN FOUR PROTEINS

Res/span Span	Mb				Ly				RNAse				Chym			
	2		3		2		3		2		3		2		3	
	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc	Obs	Calc
2	3	5.1 ± 0.3	0	0.0	3	2.3 ± 0.2	0	0.0	3	1.3 ± 0.1	0	0.0	6	4.9 ± 0.2	0	0.0
3	4	6.1 ± 0.3	0	1.4 ± 0.2	3	2.9 ± 0.3	1	0.5 ± 0.1	1	1.7 ± 0.2	1	0.2 ± 0.1	6	6.3 ± 0.3	0	1.1 ± 0.1
4	7	6.5 ± 0.4	1	2.0 ± 0.2	1	3.4 ± 0.3	2	0.7 ± 0.1	3	2.0 ± 0.3	0	0.3 ± 0.1	6	7.1 ± 0.3	2	1.6 ± 0.2
5	4	6.5 ± 0.5	2	2.5 ± 0.3	2	3.6 ± 0.4	2	0.9 ± 0.2	3	2.3 ± 0.3	0	0.4 ± 0.1	8	7.5 ± 0.4	1	2.1 ± 0.2
6	6	6.2 ± 0.5	3	2.8 ± 0.3	1	3.7 ± 0.4	1	1.2 ± 0.2	2	2.4 ± 0.3	1	0.5 ± 0.2	6	7.7 ± 0.4	3	2.6 ± 0.2
7	6	5.8 ± 0.5	2	3.1 ± 0.4	2	3.7 ± 0.4	1	1.3 ± 0.3	1	2.5 ± 0.4	2	0.7 ± 0.2	7	7.6 ± 0.5	3	2.9 ± 0.3
8	7	5.2 ± 0.5	5	3.2 ± 0.4	2	3.7 ± 0.5	3	1.5 ± 0.3	1	2.6 ± 0.4	2	0.8 ± 0.2	8	7.6 ± 0.5	1	3.0 ± 0.3
9	8	4.7 ± 0.5	2	3.3 ± 0.4	2	3.5 ± 0.5	1	1.6 ± 0.3	2	2.6 ± 0.4	1	0.9 ± 0.3	8	7.3 ± 0.5	3	3.3 ± 0.3
10	6	4.2 ± 0.5	4	3.2 ± 0.5	3	3.4 ± 0.5	1	1.7 ± 0.4	2	2.6 ± 0.5	2	1.0 ± 0.3	10	7.0 ± 0.5	3	3.5 ± 0.4

Alanine and leucine were treated as a single residue in preparing this Table. See Table III and text.

In formulating a rule pertaining to α-helical distribution it is also useful to consider the sequential distribution of the various residues in the proteins under study. Are the residues randomly distributed within a given protein and, if not, are the deviations consistent from protein to protein? A simple approach to this question is to divide each protein into successive spans of the same length and count the number of times a given residue occurs once and only once, or twice and only twice in the population of spans. The observations may be compared with that predicted for a random distribution, say, as given approximately by a Poisson distribution. In the first column of Table III the number of times alanine occurs once in regions varying in length from two to 10 is indicated, for the protein myoglobin. Thus, if myoglobin is divided into 76 spans of length two, it will be found that alanine occurs singly in 15 spans and doubly in but one span. The calculated figures are 14 ± 0.4 and 1.5 ± 0.1, respectively, where the plus-minus quantities are standard deviations. Taking into account the fact that the observed values are necessarily integral, whereas the calculated values are generally fractional, it will be seen that the observations and the calculations are in good agreement in this instance. In other instances there appear to be statistically significant deviations, as in the case of α-chymotrypsin for a span two, in which case a single alanine residue is predicted to occur 18.4 ± 0.4 times in spans of length two, but is observed only 14 times. However, and this is the main point, there does not appear to be a consistent pattern of deviations from the predicted values for any given span and for all four proteins.

The same hypothesis is tested for leucine in Table IV and again there does not appear to be a consistent pattern of deviations.

Given that individual residues are randomly distributed, at least when the proteins are considered together, it would still be possible for the distribution of pairs (or triplets, etc.) of residues to be correlated in their distribution. To test this hypothesis, alanine and leucine have been lumped together and treated as a single residue. The results are given in Table V. Again, there do not appear to be any consistent deviations.

It is concluded that the residues, whether taken singly or together, are, to a first approximation, randomly distributed. The phrase "first approximation" must be emphasized inasmuch as studies have shown that subtle correlations do, in fact, exist (28–31). However, the conclusion that the residues are randomly distributed and that, in any event, the deviations are not consistent from protein to protein, appears to be sufficiently accurate for the purposes of the following model. Alternatively, Tables III–V support the hypothesis that, given deviations from random, these are not consistent from protein to protein.

MODEL

The proposed model is based on the assumption that there are two, three, or possibly four classes of residues: (*a*) primary helix formers (PHF), (*b*) secondary helix formers (SHF), (*c*) indifferent residues (IR), and (*d*) helix destabilizers (HD). By

hypothesis, when the primary helix formers (PHF) exceed a certain threshold (see Appendix for definitions), a helix will be "nucleated" in that segment of the polypeptide chain where the threshold is exceeded. Furthermore, it is assumed that once nucleation has occurred, growth can take place either from one end or both ends provided some second, presumably lower, threshold is exceeded. Growth is assumed to occur if either PHF or secondary helix formers (SHF), or both exceed this threshold level. It is convenient to think of growth as taking place in increments of some given length (i.e., given number of residues). It is also useful to think of cycles of growth. In the first cycle the first increment is examined to see if the thresh-

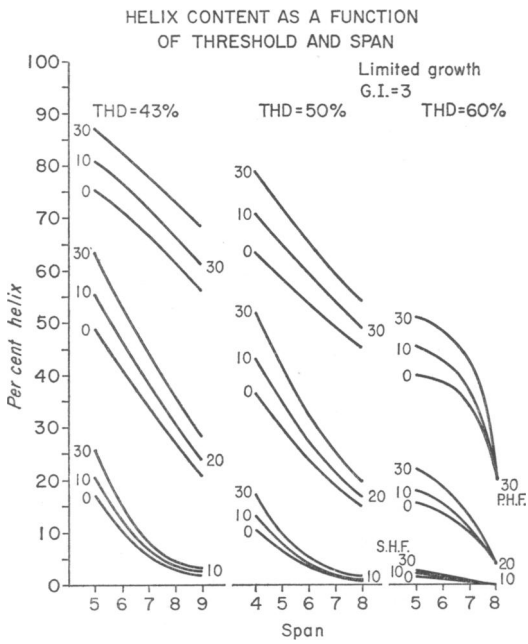

FIGURE 4 The averaged helical content of 100 proteins (hypothetical) containing varying percentages of primary (PHF) and secondary (SHF) helix formers. Thus the curves on the right correspond to the average helical content when the PHF vary over the range of 10, 20, and 30% and the SHF vary over the range 0, 10, and 30%. One cycle of growth with increment (GI) of three.

old is exceeded, and if it is, then the next increment is examined, and so on. The characteristics associated with no growth (NG), one cycle of growth (limited growth), and repeated cycles of growth (unlimited growth) will be examined. The values of thresholds (THD), spans, per cent helix formers, and growth increments (GI) to be considered are summarized below.

Average threshold per cent	Span No. residues	Helix formers per cent		Growth increment No. residues
		PHF	SHF	
43	5, 7, 9			
50	4, 6, 8	10, 20, 30	0, 10, 30	3, 4
60	5, 7, 8			

An experiment is carried out as follows. A given number of PHF (say 10) is randomly inserted in a protein 100 residues long (i.e., using a computer program). Then a given number of SHF (say 10) are also inserted, again randomly. The protein is now allowed to nucleate in those regions where the threshold (say three PHF in a span of six, i.e. 50%) is exceeded. The nucleated regions are noted (this corresponds to no growth). Then one cycle of growth is initiated, say with a growth increment (GI) of three. Next, repeated cycles of growth are carried out until no further change occurs (i.e. unlimited growth) and again the results are noted. The total helical content is calculated for the three cases of no growth (nucleation only), limited

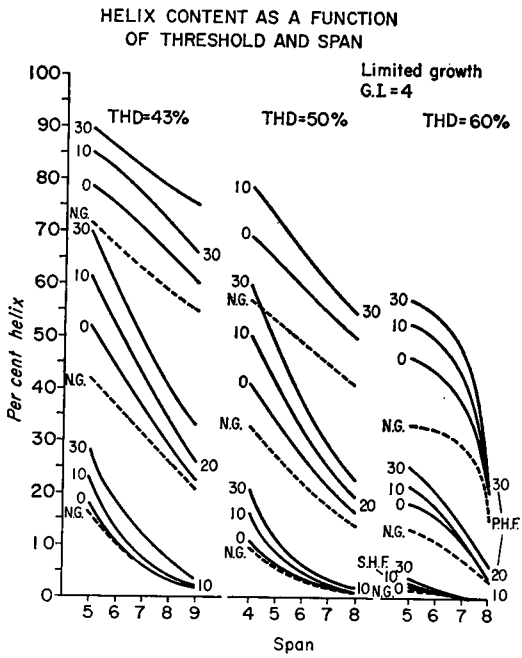

FIGURE 5 See legend for Fig. 4. Differs from Fig. 4 in that growth increment (GI) is four. Results for no growth (NG) are indicated by broken line.

growth, and unlimited growth. Finally, the whole experiment is repeated 100 times and the mean helical content is calculated. Similarly, 100 experiments were carried out for all combinations of the values given above. The first two residues were not included in computing the helical regions (see below). A growth threshold of 25% (i.e. one in four) was used whenever the growth increment was four and a threshold of 33% whenever the growth increment was three.

The results of the experiments are summarized in Figs. 4, 5, and 6. The results in Fig. 4 correspond to limited growth, with a growth increment of three. It will be seen that raising the threshold, especially from 50 to 60%, dramatically decreases the helical content. On the other hand, introducing SHF generally has a much smaller effect. Increasing the span at constant threshold also decreases the α-helical content, in some cases markedly.

Fig. 5 gives the results for limited growth, with a growth increment of four, as well as the results for no growth (nucleation only). It will be seen, on comparing Figs. 4 and 5, that increasing the growth increment from three to four has a very small effect, as would be expected. Except at the higher threshold, going from the condition of no growth to growth without SHF also has little effect.

Fig. 6 gives the results for unlimited growth, with a growth increment of four. For high levels of PHF and SHF the helical content becomes essentially independent of span at the lower thresholds. The helical content is more sensitive to SHF content than it is in the case of limited growth.

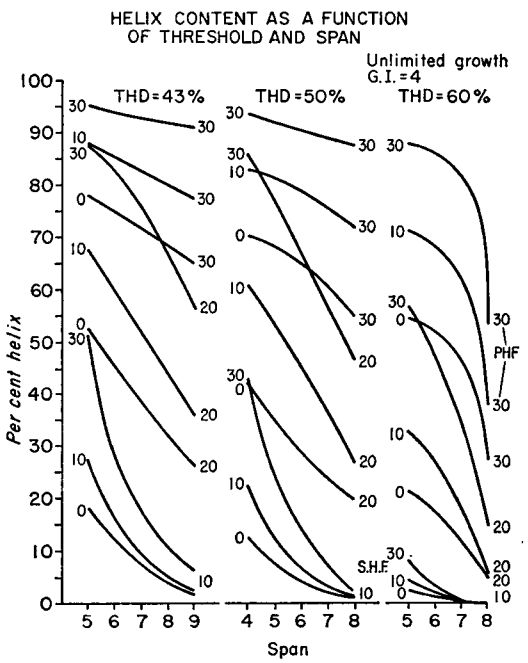

FIGURE 6 See legend for Fig. 4. Differs from Fig. 4 in that growth increment (GI) is four and growth is unlimited.

It should be noted that the thresholds indicated in the above figures are mainly average values. In the case when the threshold is 43%, the actual values are 40% ($\frac{2}{5}$), 43% ($\frac{3}{7}$), and 44% ($\frac{4}{9}$). Again, where the threshold is 60%, the actual values are 60% ($\frac{3}{5}$), 57% ($\frac{4}{7}$), and 62% ($\frac{5}{8}$).

APPLICATION OF MODEL

In order to apply the model to proteins it is necessary to make some assumption as to which residues may be PHF. On the basis of Table II it will be assumed for the moment, that alanine and leucine are the PHF and, in fact, the only PHF. It is then possible to compare helical content expressed as a function of alanine plus leucine content, as seen in Fig. 3, with that which would be predicted by the above model for the same percentage of PHF. We are particularly interested now in the effects of

varying threshold at constant span. One can study these effects by interpolating in Figs. 4–6 to read off the helical content for any given span. Thus, although a threshold of 50% is not particularly meaningful for a span of five, say, in the present model (i.e. 2½ PHF), nevertheless, one can, by interpolation in Figs. 4–6 determine what the helical content would be if the threshold were 50%.

The data have been plotted in Fig. 7 for the case of unlimited growth with a growth increment of four and a nucleation span of five. It will be seen that the agreement between the model (solid lines) and the observations (dotted lines) is about equally poor for all three thresholds. It is notable that the per cent helix predicted by the

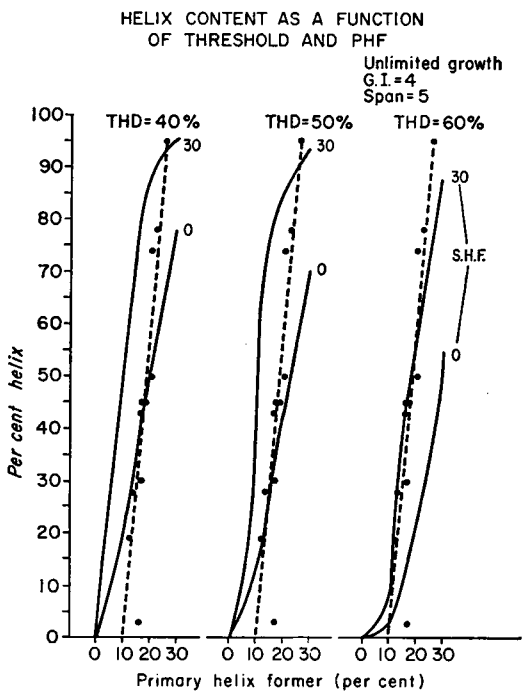

FIGURE 7 The smooth curves are derived from Fig. 6 by interpolation. Effectively they indicate how the helical content varies with the per cent of PHF (i.e., span constant). The dotted line represents the least squares regression line relating helical content in 11 proteins to the alanine plus leucine content (cf. Fig. 3).

model is very sensitive to the SHF content. If SHF do exist, and vary in amount from protein to protein in the range 0–30% as would seem reasonable, then unlimited growth is perhaps unlikely. That is, the close correlation between helical content and alanine plus leucine content would necessarily be spurious.

The data are plotted in Fig. 8 for the case of limited growth, with a growth increment of three and a nucleation span of four. It will be seen that the fit is much better at 43 and 50% than at 60%.

In Fig. 9 the data are plotted for the same case as Fig. 8, but with a span of five. Also, the data are plotted for the case of no growth. It will be seen that the fit is best

at the lowest threshold (40%, i.e. $\frac{2}{5}$). As might be expected, the difference between no growth and growth with no SHF is very slight.

Finally, Fig. 10 shows the effect of increasing the span to six. The fit at the lowest threshold is now quite good. Furthermore, the helical content is only weakly dependent upon the SHF content.

The fact that a good fit can be obtained between the model and the observations is an interesting result (i.e., Fig. 10). Furthermore, the fact that the model is not, in the given range (THD = 43%, span = 6), very sensitive to SHF content is en-

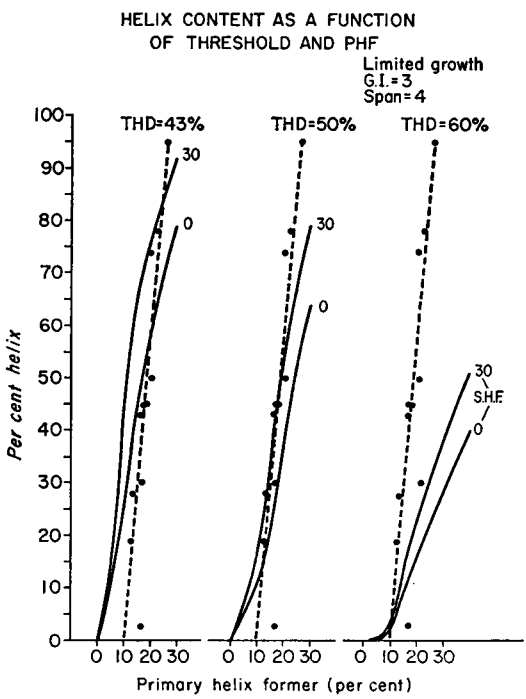

FIGURE 8 See legend for Fig. 7. Differs from Fig. 7 in that growth is limited, the growth increment (GI) is only three and the span is only four. Derived from Fig. 4, in part by extrapolation.

couraging because it gives hope that the helical content can be predicted tolerably well just from a knowledge of the PHF content.

It had been shown earlier without justification that a crude rule based on this type of argument would work, at least in the case of myoglobin and lysozyme (35). Since that time, two further structures have become available and it is accordingly of interest to see if a somewhat refined rule, based on this model, is applicable both to the old and the new data.

In formulating a rule designed to predict helical segments of a polypeptide, it is essential to employ some measure of "goodness of fit," inasmuch as visual estimates of goodness of fit are subjective and possibly misleading. Two measures of goodness

of fit (GF) will be employed:[1]

$$GF_1 = \left(\frac{N_T - N_I}{N_T}\right) \times 100 \qquad (1)$$

$$GF_2 = \left(1 - \left|\frac{N_{HO} - N_{HC}}{N_{HO}}\right|\right) \times 100 \qquad (2)$$

Where N_T is the total number of residues, N_I is the number of residues predicted incorrectly (i.e. predicted to be helical or nonhelical when the reverse is the case),

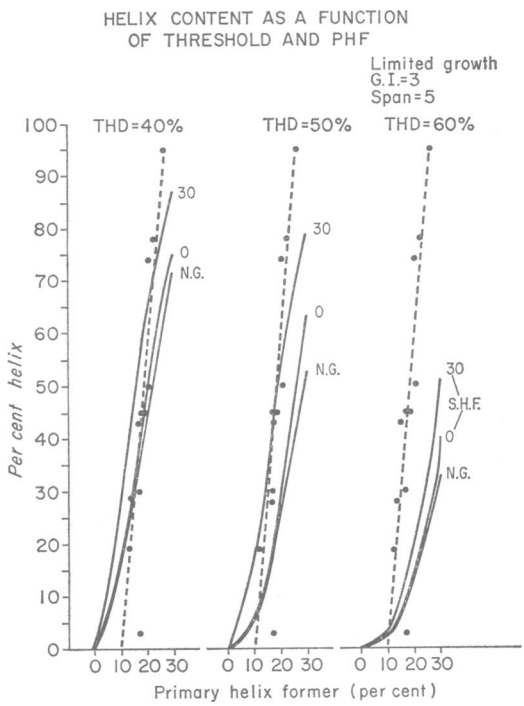

FIGURE 9 See legend for Fig. 7. Same as Fig. 8 but for span, which is increased to five. Results are shown for no growth (NG).

N_{HO} is the observed number of helical segments, and N_{HC} is the calculated number of helical segments (i.e. neglecting abrupt corners). For a given rule and a given protein, the two measures of goodness of fit will be written in the form GF_1/GF_2.

[1] A more stringent measure of goodness of fit is obtained by subtracting the number wrong (N_I) from the number right. In the above nomenclature this is equivalent to:

$$GF = \left(\frac{N_T - 2N_I}{N_T}\right) \times 100.$$

This measure is 50% when GF_1 is 75% (see Table VI). It is a matter of judgment whether the more pessimistic estimate is more realistic. Less stringent measures than GF_1 are also possible.

Note that whereas the first measure (GF_1) is a reliable indication of one type of goodness of fit (i.e. net agreement), the second is merely a measure of the number of discontinuities in the chain introduced by the rule. For example, a rule with a short span (say three) and an intermediate or low value of threshold would be expected to give rise to many short helical regions. GF_2 is a measure of this tendency to produce discontinuities calculated without regard to whether or not the helical regions are correct. Also, note that GF_2 is 100% when the number of predicted helical regions is correct, but is less than 100% if the predicted number is either less than or greater than the observed one. Finally, it may be observed that if GF_1 is

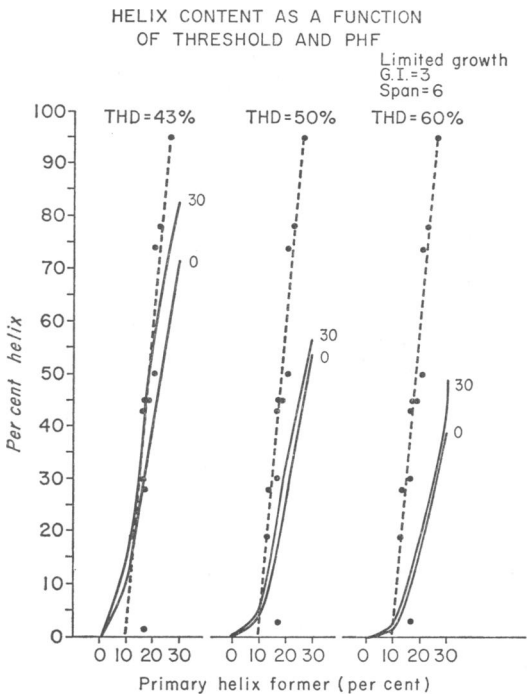

FIGURE 10 See legend for Fig. 7. Same as Figs. 8 and 9 but for span which is increased to six.

high (say 75%), then GF_2 is a fair measure of the number of helical segments correctly predicted. The converse is not true.

The two measures of goodness of fit have been calculated for six rules (see Table VI) as applied to the seven proteins given in Table I.

Rule I. If the sum of alanine plus leucine exceeds 20%, the protein is 100% helical; whereas if the sum is less than 20%, the protein is 0% helical. This may be termed the "all-or-nothing" rule. GF_2 has not been calculated for this case inasmuch as the fit is necessarily very poor. As determined by GF_1, however, the rule is quite reasonable. This rule may be regarded as a limiting case in which the residues are evenly distributed throughout the protein (i.e. highly ordered protein) and the threshold is 20%. In this event the helical content must be either 0 or 100%.

Rule II. The threshold is 40%, the span five, the PHF are alanine, leucine, and glutamate (see below), and there is no growth. In calculating mean 1, the value of GF_2 for α-chymotrypsin has not been included since it would dominate the expression. Suffice to say that Rule II and the subsequent rules predict too many helical regions for α-chymotrypsin. In this and the following rules, the first two residues have been treated as nonhelical.

Rule III. The threshold is 43% and the span seven. Otherwise, the rule is the same as Rule II. Agreement is slightly better than for Rule II.

Rule IV. The threshold is 44% and the span nine. Otherwise, the rule is the same as Rule II. There is a marked drop in the mean values of GF_1 and GF_2 (see mean 1) as well as in the mean of GF_1 and GF_2 taken together (i.e., mean 2).

TABLE VI

Protein	No. helical regions observed	Goodness of fit (per cent)				
Rule	—	1	2	3	4	5
Thr/span	—	—	2/5	3/7	4/9	3/7
GI	—	—	0	0	0	3
Mb	6	78	77/50	77/100	64/71	67/83
Ly	7	57	76/57	73/43	68/29	65/14
RNAse	3	81	73/66	86/66	89/66	86/66
Hb-α	7	77	70/71	57/71	58/57	52/57
Hb-β	7	78	75/100	71/86	58/57	62/71
Hb-γ	7	78	66/100	65/86	36/43	45/57
Chym	1	97	61/−900	73/−500	90/0	92/0
Mean 1	—	78	71.1/74.0	71.7/75.3	66.1/53.8	67/58
Mean 2	—	—	72.6	73.5	60.0	62.5

The table summarizes the goodness of fit obtained in predicting the helical regions of the proteins myoglobin (Mb), lysozyme (Ly), ribonuclease (RNAse), the α, β, and γ-hemoglobins (Hb-α, Hb-β, Hb-γ) and α-chymotrypsin by rules 1–5. See text.

Rule V. This rule is identical to Rule III but for the fact that limited growth, with a growth increment of three, is allowed. Again, there is a marked drop in the goodness of fit as compared to Rules I, II, and III.

Rule VI. The threshold is 60%, the span is five, there is no growth, and the PHF are alanine, leucine, and glutamate. This may be termed a "safe" rule, in the sense that the rule is designed to predict reliably only a proportion of the α-helical content of proteins. This result is to be achieved by raising the threshold. The results are summarized separately in Table VII. The first column gives the number of residues predicted to be helical and the third the number which are correct. The second gives the observed number. In the next to last column the ratio of the number of residues correctly predicted to the total number of helical residues expressed as a percentage is given. On the average, about one-third of the helical residues are correctly pre-

TABLE VII

Protein	No. helical residues predicted (*a*)	No. helical residues observed (*b*)	No. residues corr. predicted (*c*)	Per cent (*c*/*b*)	Per cent (*c*/*a*)
Mb	54	119	46	39	85
Ly	21	56	20	36	95
RNAse	13	24	8	33	62
Hb-α	44	108	33	31	75
Hb-β	47	113	37	33	79
Hb-γ	33	113	30	27	91
Chym	29	8	6	75	21

The number of helical residues predicted by the rule that any region of five residues which includes at least three of the residues alanine, leucine, or glutanate will be helical. The first "per cent column" indicates the proportion of helical residues predicted, and the second "per cent column" indicates the accuracy of the prediction. Proteins are myoglobin (Mb), lysozyme (Ly), the α, β, and γ-hemoglobins (Hb-α, Hb-β, Hb-γ) and α-chymotrypsin. This is intended to be a "safe" rule which predicts a small proportion of the helical content reliably.

dicted. In the last column the residues which were correctly predicted are expressed as a percentage of all the residues predicted to be α-helical. Even including α-chymotrypsin, which has only eight helical residues and for which agreement is poor, the rule is successful on the average in predicting about one-third of the helical residues with about 75% reliability (i.e., three out of four residues which are predicted to be helical are predicted correctly).

DISCUSSION

It has been shown that the alanine plus leucine content of 11 proteins (see Fig. 3) is well correlated with helical content. It has further been shown that this correlation is explicable if alanine and leucine act as primary helix formers to nucleate α-helical regions in a polypeptide chain. A rule based on a nucleation model with randomly distributed helix formers does give a reasonable fit (70–75%) with the known protein structures. Glutamate was included as a PHF on the basis of the analysis summarized on Table II. Since glutamate and glutamine are not distinguished generally in the amino acid composition tables, a plot of alanine plus leucine plus glutamate (cf. Fig. 3) can only be made for four distinct proteins (myoglobin, lysozyme, ribonuclease, and α-chymotrypsin). The effect, in Fig. 10 for example, of including glutamate would be to move the dotted line somewhat to the right and to reduce the slope. It is possible that the agreement would be improved, but until more data are available, it is preferable to compare the data for alanine plus leucine in 11 proteins to the model rather than alanine plus leucine plus glutamate in four proteins. Valine, which was formerly regarded as a PHF (35), has been omitted because of its inconsistent ordering in Table II.

Some caution is warranted in concluding that alanine, leucine, and glutamate,

for example, are in fact PHF. Until more structures are examined, it cannot be ascertained whether a correlation between the distribution of one or more residues and the distribution of α-helices is due to the given residues or to the displacement of other, possibly destabilizing, residues. More complex effects are also possible.

Nevertheless, a simple rule which attributes the nucleation of α-helical regions to alanine, leucine, and glutamate is in reasonable agreement with the observations. Until more data are available, attempts to obtain a better fit by taking into account detailed effects such as those of proline, or a vector type of nucleation and growth, or specific interactions between residues, do not appear to be warranted.

APPENDIX

Definition of Terms

Primary helix formers (*PHF*) are those residues whose presence at threshold level is necessary and sufficient to produce α-helix nucleation.

Secondary helix formers (*SHF*) are those residues whose presence at threshold level is sufficient to produce growth of an α-helix.

Helix destabilizers are those residues whose presence at threshold level is sufficient to prevent the nucleation and/or growth of α-helices.

Indifferent residues are those residues which are neither helix formers nor destabilizers.

Span refers to a segment of given length of a polypeptide chain (i.e., a given number of consecutive residues).

Threshold is the number of helix formers (i.e., PHF or SHF) required to produce either nucleation or growth, expressed as a percentage of either the span or growth increment. Thus, if three residues (PHF) in a span of six are required for nucleation, then the nucleation threshold is 50%. If one residue (PHF or SHF) in an increment of three is required for growth then the growth threshold is 33%.

Growth increment (*GI*) is the number of residues considered to be involved in one cycle of growth. Thus if the growth increment is three, growth will be considered to take place by the addition of three residues to the helical segment in each cycle of growth.

Per cent helix is derived by dividing the number of residues in helical segments by the total number of residues and multiplying by 100.

Limited growth is one cycle of growth after nucleation.

Unlimited growth is repeated cycles of growth carried out until growth ceases.

Received for publication 13 June 1968.

REFERENCES

1. EDMUNDSON, A. B. 1965. *Nature.* **205**:883.
2. KENDREW, J. C., R. E. DICKERSON, B. E. STRANDBERG, R. G. HART, D. R. DAVIES, D. C. PHILLIPS, and V. C. SHORE. 1960. *Nature.* **185**:422.
3. CANFIELD, R. 1963. *J. Biol. Chem.* **238**:2698.
4. BLAKE, C. C. F., D. F. KOENIG, G. A. MAIR, A. C. T. NORTH, D. C. PHILLIPS, and V. R. SARMA. 1965. *Nature.* **206**:757.
5. SMYTH, D. G., W. H. STEIN, and S. MOORE. 1963. *J. Biol. Chem.* **238**:227.
6. KARTHA, G., J. BELLO, and D. HARKER. 1967. *Nature.* **213**:862.

7. HARTLEY, B. S., and D. L. KAUFFMAN. 1966. *Biochem. J.* **101**:229.
8. MATTHEWS, B. W., P. B. SIGLER, R. HENDERSON, and D. M. BLOW. 1967. *Nature.* **214**:652.
9. BRAUNITZER, G., R. GEHRING-MÜLLER, N. HILSCHMANN, K. HILSE, G. HOBOM, V. RUDLOFF, and B. WITTMANN-LIEBOLD. 1961. *Z. Physiol. Chem.* **325**:283.
10. KONIGSBERG, W., G. GUIDOTTI, and R. H. HILL. 1961. *J. Biol. Chem.* **236**:P.C. 55.
11. SCHROEDER, W. A., J. R. SHELTON, J. B. SHELTON, and J. CORNICK. 1963. *Biochemistry.* **2**:1353.
12. BLOUT, E. R., C. DE LOZE, S. M. BLOOM, and G. D. FASMAN. 1960. *J. Am. Chem. Soc.* **82**:3787.
13. BLOCK, H., and J. A. KAY. 1967. *Biopolymers.* **5**:243.
14. BRADY, G. W., and R. SALOVEY. 1967. *Biopolymers.* **5**:331.
15. FRAZER, R. D. B., B. S. HARRAP, R. LEDGER, T. P. MACRAE, F. H. C. STEWART, and E. SUZUKI. 1967. *Biopolymers.* **5**:797.
16. ZIMM, G. H., and J. K. BRAGG. 1959. *J. Chem. Phys.* **31**:526.
17. SCHWARZ, G. 1967. *Biopolymers.* **5**:321.
18. RAMACHANDRAN, G. N., C. RAMAKRISHNAN, and V. SAISEKHARAN. 1963. *In* Aspects of Protein Structure. G. N. RAMACHANDER, editor. Academic Press Inc., New York. 121.
19. NEMETHY, G., and H. A. SCHERAGA. 1965. *Biopolymers.* **3**:155.
20. MILLER, W. G., and P. J. FLORY. 1966. *J. Mol. Biol.* **15**:298.
21. CRAIG, M. E., and D. M. CROTHERS. 1968. *Biopolymers.* **6**:385.
22. GO, N., M. GO, and H. A. SCHERAGA. 1968. *Proc. Natl. Acad. Sci.* **59**:1030.
23. DAVIES, D. R. 1964. *J. Mol. Biol.* **9**:605.
24. GUZZO, A. V. 1965. *Biophys. J.* **5**:809.
25. PERUTZ, M. F., J. C. KENDREW, and H. C. WATSON. 1965. *J. Mol. Biol.* **13**:669.
26. SCHIFFER, M., and A. B. EDMUNDSON. 1967. *Biophys. J.* **7**:121.
27. COOK, D. A. 1967. *J. Mol. Biol.* **29**:167.
28. BIGELOW, C. 1967. *J. Theoret. Biol.* **16**:187.
29. THIEBAUX, H. J., and H. H. PATTEE. 1967. *J. Theoret. Biol.* **17**:121.
30. KRZWICKI, A., and P. P. SLONIMSKI. 1967. *J. Theoret. Biol.* **17**:136.
31. PERITI, P. F., G. QUAGLIAROTTI, and A. M. LIQUORI. In press.
32. DUNNILL, P. 1968. *Biophys. J.* **8**:865.
33. ISING, E. 1925. *Z. Physik.* **31**:253.
34. MOROWITZ, H. J. 1964. Cited by G. H. Haggis, D. Michie, A. R. Muir, K. B. Roberts, and P. M. B. Walker. *In* Introduction to Molecular Biology. John Wiley & Sons, Inc., New York. 43.
35. PROTHERO, J. W. 1966. *Biophys. J.* **6**:367.