THE INFLUENCE OF AMINO-ACID SEQUENCE ON PROTEIN STRUCTURE

ANTHONY V. GUZZO

From the Department of Biophysics, University of Chicago, Chicago. Dr. Guzzo's present address is the Department of Chemistry, The University of Wyoming, Laramie.

ABSTRACT On the basis of the known sequences and structures of myoglobin, and alpha and beta hemoglobin, a possible correlation between certain amino acids in the sequence and the location of the helical and non-helical parts of the structure is suggested. The presence in the sequence of four critical groups; proline, aspartic acid, glutamic acid, or histidine appears to be necessary (although the last three are not sufficient) for a helical disruption to form. Additional support for this correlation is obtained from analyses of proline replacement in mutant and variant proteins. A mechanism based on hydrophobic bonding is proposed as a rationale for the apparent behavior of these groups. On the basis of these rules and correlations, secondary structures can be proposed for lysozyme and tobacco mosaic virus protein which are consistent with several pieces of evidence.

INTRODUCTION

The present note is concerned with an explanation of some apparent correlations between protein amino acid sequence and structure, and a suggestion for a physical-chemical rationale for these correlations.

There is increasing experimental evidence in favor of the hypothesis that the information necessary to construct the enzymatically active secondary and tertiary structure of a protein in a given solvent may be determined entirely by the sequence of amino acids; or briefly, that "sequence implies structure." That is, the interactions between the amino acids and with the solvent may suffice to determine the structure without the need of any further information or interaction with other components of the cell. Epstein *et al.* (1963) have summarized this evidence which shows that by careful control of the experimental conditions, proteins can be reversibly denatured and renatured. Even in those cases where cystine cross-links have been cleaved by reduction, such as in Anfinsen's study of ribonuclease (Anfinsen, 1961) and Levinthal's study of alkaline phosphatase (Levinthal, 1962), renaturation by careful oxidation leads to a recovery of 80 to 90 per cent of the original enzyme activity.

In general then, these results imply that at least for these molecules the thermodynamically most stable form in aqueous solution is the enzymatically active, native configuration.

However, the prediction of what this form will be for a given sequence presents a considerable problem. There are a number of reasons why it is not possible to calculate *a priori* the most stable structure, even if one assumes only classical interactions; *i.e.,* ion-dipole, dipole-dipole interactions, etc. Solvent interactions, including hydrophobic bonding effects (Kauzmann, 1959) are extremely important yet the exact manner in which they should be included is not well understood. The contribution to the stability of the structure by hydrogen bonding is not known since in the presence of water—an excellent hydrogen-bonding agent—the carbonyl-amide bonds are expected to be quite weak, yet they may be considerably stronger in the interior of the molecule or in other anhydrous regions (Nemethy *et al.,* 1963).

It therefore becomes important to inquire if there are any empirical rules or correlations that could reduce the problem into parts that might be more easily understood separately. One possible working hypothesis is that it is approximately correct to discuss protein secondary structure—the division of the amino acid sequence into helical and non-helical parts—apart from the tertiary structure, which includes the folding of these regions into a three dimensional form (we will make the common assumption that the alpha helix is the predominant helical form). This is equivalent to postulating that the interaction energies involved in maintaining the secondary structure are, in general, greater than those involved in forming the tertiary structure. Thus if this hypothesis holds, the influence of the amino acid sequence may be discernible in the secondary structure of a given protein in such a way that helical breaks or non-helical regions may tend to be associated regularly with certain amino acids or combinations of them.

Sequence-Structure Correlations. Blout (1962) has pointed out that the known polyamino acids fall into two categories depending upon their ability to form alpha helical structures in solution. Those polyamino acids which do not exist in alpha helical form are polymers of valine, isoleucine, proline, serine, threonine, and S-methyl-cysteine. It was assumed that steric factors are responsible for the non-helical nature of poly-L-proline, poly-L-isoleucine, and poly-L-valine, and that similar factors may contribute to the weakening of the helical structure in proteins whenever these groups occur in the sequence. In addition, polymers of serine, threonine, and S-methyl cysteine do not occur as alpha helices for reasons possibly related to the proximity of the heteroatom to the peptide backbone.

Davies (1964) has shown that an inverse correlation exists between the concentrations of the above mentioned six amino acids, in many proteins, with the helix content of these proteins as measured by optical rotatory dispersion, that is, the greater the total concentration of these groups, the less the helical content.

Thus this direction of thinking has seemed to be worth pursuing to see if there

is any correlation between these or any other amino acids and the non-helical regions, in proteins whose sequence and structure are known. Kendrew's determination of the myoglobin structure (Kendrew *et al.*, 1960) along with Edmundson's sequence data (Edmundson, 1965) provided the starting point. In addition, extensive use was made of Perutz's structure data (Perutz *et al.*, 1960) and Braunitzer's sequence data on the hemoglobins (Braunitzer *et al.*, 1964). Data necessary for the discussion have been reproduced in Table I.

RESULTS

Table II shows the statistical distribution of the various amino acids in the helical and non-helical sections of the proteins. For the present purpose the helical sections were defined so that they consisted of only those amino acids which are hydrogen bonded on both the amide and carbonyl groups of the peptide linkage. Thus, four amino acids on each end of each helical section as given by Kendrew were considered to belong to the non-helical regions. This seems reasonable since any amino acids that have a helix-disrupting function may be located at the ends of helices (as proline frequently is) rather than entirely outside in the non-helical region. This convention leads to a total "non-helical" content of 64 per cent for the four proteins studied, so that the numbers obtained were normalized by dividing by 64, while the number in the helical regions were normalized by dividing by 36. The ratio of these normalized numbers is given in column 4 of Table II; when it is significantly greater than 1 it indicates a tendency for the amino acid to be associated with the non-helical region. Column 5 shows how the ratios are changed if a slightly different definition of the helical regions is used; if only the last *three* amino acids from each helix end are assigned to the non-helical regions, the tendency of certain amino acids to accumulate in these regions is not substantially altered. With either definition, aspartic acid in particular is seen to prefer the non-helical regions by a factor of three. This is quite far removed from what would be expected if its distribution were random.

Upon examination of Table II in light of Blout's work, there does not appear to be any correlation of the presence of valine, isoleucine, serine, and threonine with the non-helical regions. The importance of cysteine could not be determined since it appears so infrequently in these proteins. Proline, on the other hand, always appears at bends or non-helical regions, as is well known. In addition, a strong correlation seems to exist between non-helical nature and the presence of the charged amino acids; aspartic acid, glutamic acid, and histidine. These groups, which we have referred to as "critical groups," have been reproduced in boldface type in Table I. Moreover, these critical groups appear repeatedly at nearly all of the non-helical regions or bends. Thus, at least for these proteins examined, aspartic acid, glutamic acid, and histidine appear to have an intimate relationship to the secondary structure and one might postulate that the presence of one of these amino

TABLE I

The sequences and secondary structural outlines of human alpha, beta, and gamma hemoglobin and myoglobin. Adapted from Table XI of Braunitzer *et al.* (1964). The critical groups have been reproduced in **boldface type.** The sequence of myoglobin has been revised according to Edmundson (1965) with the initial section aligned with the hemoglobin sequences as well as possible. The helical parts are indicated by the dashed line connecting the letters as given by Kendrew (1964).

α Val- -Leu-Ser-Pro-Ala-Asp-Lys-Thr-Asn-Val-Lys-Ala-Ala-Try-Gly-Lys-Val-Gly-Ala-
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

β Val-His-Leu-Thr-Pro-Glu-Glu-Lys-Ser-Ala-Val-Thr-Ala-Leu-Try-Gly-Lys-Val-Asn-
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

γ Gly-His-Phe-Thr-Glu-Glu-Asp-Lys-Ala-Thr-Ile-Thr-Ser-Leu-Try-Gly-Lys-Val-Asn-
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

M -Val-Leu-Ser-Glu-Gly-Glu-Try-Gln-Leu-Val-Leu-His-Val-Try-Ala-Lys-Val-Glu-Ala-
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

A - A
← Helix →

α His-Ala-Gly-Glu-Tyr-Gly-Ala-Glu-Ala-Leu-Glu-Arg-Met-Phe-Leu-Ser-Phe-Pro-Thr-Thr-
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

β -Val-Asp-Glu-Val-Gly-Gly-Glu-Ala-Leu-Gly-Arg-Leu-Leu-Val-Val-Tyr-Pro-Try-Thr-
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

γ -Val-Glu-Asp-Ala-Gly-Gly-Glu-Thr-Leu-Gly-Arg-Leu-Leu-Val-Val-Tyr-Pro-Try-Thr-
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38

M Asp-Val-Ala-Gly-His-Gly-Gln-Asp-Ile-Leu-Ile-Arg-Leu-Phe-Lys-Ser-His-Pro-Glu-Thr-
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

B - B C - - - - - - -

α Lys-Thr-Tyr-Phe-Pro-His-Phe- -Asp-Leu-Ser-His- Gly-Ser-Ala-
40 41 42 43 44 45 46 47 48 49 50 51 52 53

β Gln-Arg-Phe-Phe-Glu-Ser-Phe-Gly-Asp-Leu-Ser-Thr-Pro-Asp-Ala-Val-Met-Gly-Asn-Pro-
39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

γ Gln-Arg-Phe-Phe-Asp-Ser-Phe-Gly-Asn-Leu-Ser-Ser-Ala-Ser-Ala-Ile-Met-Gly-Asn-Pro-
39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58

M Leu-Glu-Lys-Phe-Asp-Arg-Phe-Lys-His-Leu-Lys-Thr-Glu-Ala-Glu-Met-Lys-Ala-Ser-Glu-
40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59

- - - - - - C D - - - - - - - - - - - - - - - D E - - - - -

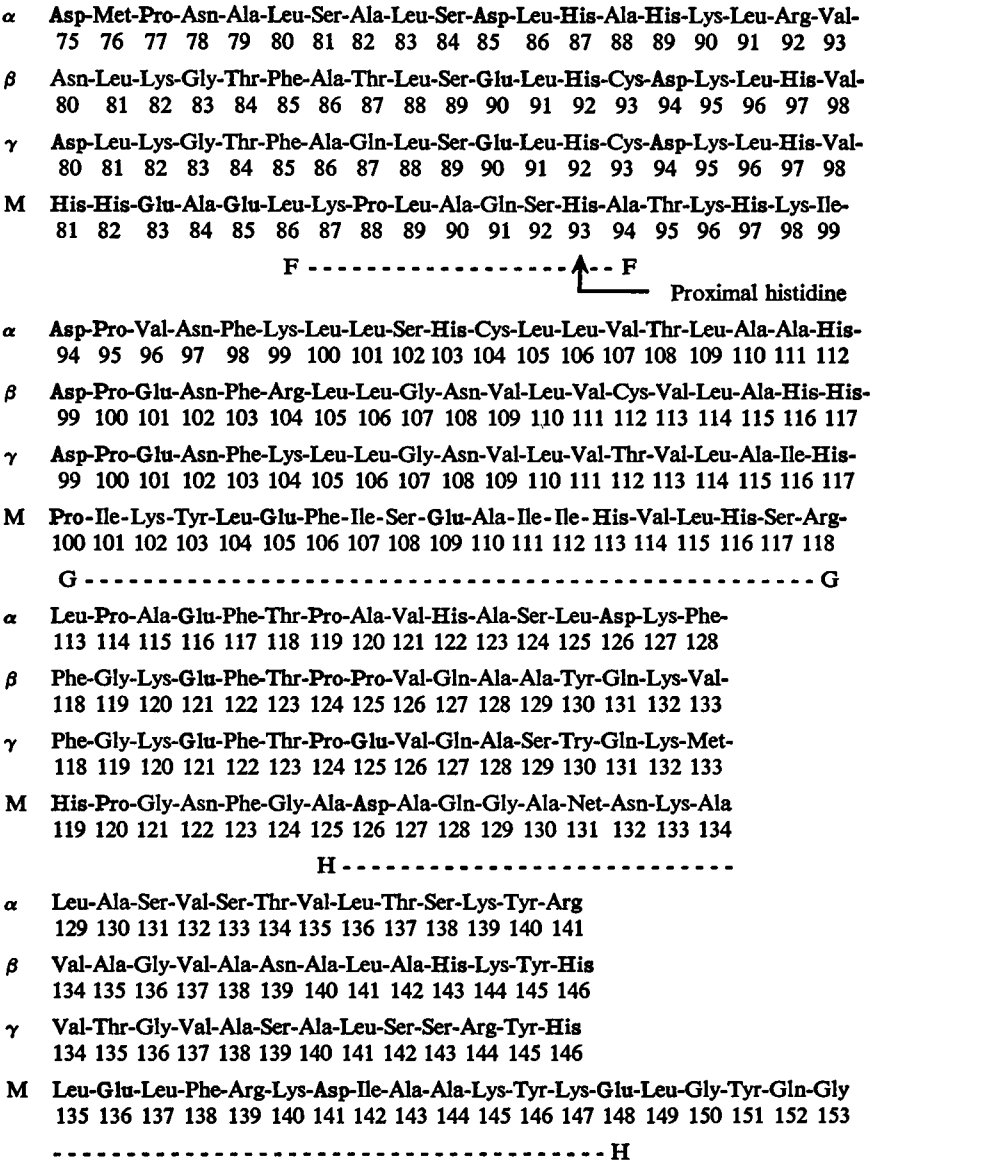
α Gln-Val-Lys-Gly-His-Gly-Lys-Lys-Val-Ala-Asp-Ala-Leu-Thr-Asn-Ala-Val-Ala-His-Val-Asp-
54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74

β Lys-Val-Lys-Ala-His-Gly-Lys-Lys-Val-Leu-Gly-Ala-Phe-Ser-Asp-Gly-Leu-Ala-His-Leu-Asp-
59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79

γ Lys-Val-Lys-Ala-His-Gly-Lys-Lys-Val-Leu-Thr-Ser-Leu-Gly-Asp-Ala-Ile-Lys-His-Leu-Asp-
59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79

M Asp-Leu-Lys-Lys-His-Gly-Val-Thr-Val-Leu-Thr-Ala-Leu-Gly-Ala-Ile-Leu-Lys-Lys-Lys-Gly-
60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80

- - - - - - - - - - - ↑ - E
└── Distal histidine



acids in a section of a protein might be a *necessary, though not sufficient, condition* for the section to be non-helical.

The exceptions to this suggested relationship are themselves interesting. In particular, the region 45 to 50 in gamma hemoglobin does not contain any critical groups, yet it is disordered; in this case the 1,2,4 spacing of the serines 49, 50, and 52 may have some disruptive effect on the helix, but it is not clear how. In the

TABLE II

DISTRIBUTION OF AMINO ACIDS IN HELICAL AND NON-HELICAL SECTIONS OF THE PROTEINS GIVEN IN TABLE I

1 Amino acid	2 Number out of helix*	3 Number inside helix	4 Normalized ratio*	5 Normalized ratio‡
Asp	22	4	3.1	3.5
Tyr	9	2	2.5	2.1
Glu	26	7	2.1	1.4
His	29	8	2.0	1.7
Phe	22	7	1.7	2.5
Val	28	24	0.7	0.7
Ile	6	7	0.5	0.7
Thr	16	13	0.7	0.6
Ser	18	14	0.7	0.7
Pro	19	0	–	–
Lys	31	18	0.8	1.1
Arg	7	6	0.6	1.0
Try	2	6	0.2	0.3
Ala	32	29	0.6	0.7
Leu	39	29	0.8	0.5
Gly	22	21	0.6	0.7
Met	5	2	1.6	1.1
Cys	2	2	0.6	0.8
Gln + Asn	21	10	1.2	0.6

*"out of helix" includes 4 amino acids at the ends of the helical sections as given by Kendrew.
‡Calculated including only 3 amino acids at the ends of the helical sections.

hemoglobins, there are no critical groups present to account for the bend at the beginning of the F helix. This may be related to the binding of the heme itself, which is at histidine 87 in the alpha chain.

Table II suggests that phenylalanine and tyrosine might also have a correlation with the non-helical regions, although this is statistically less reliable in the case of tyrosine. This appearance, at least in the case of phenylalanine, is largely due to its high concentration in the two non-helical regions C to D and G to H of the proteins of Table I. Phenylalanine does not seem to be concentrated preferentially in non-helical regions elsewhere.

Additional Experimental Support. Until a greater number of x-ray determinations become available, experimental support for the suggested unique behavior of aspartic acid, glutamic acid, and histidine can be only indirect. However, in addition to the general correlation of these groups with the non-helical features of proteins as mentioned above, a more specific kind of evidence may be obtained from considering cases of "proline substitution." If proline is replaced by another amino acid in a mutant or variant protein and the molecular species remains viable, it seems reasonable to suppose that the protein secondary and tertiary structures have

not been greatly changed. (Viability would seem to be a more stringent requirement to insure tertiary structure preservation upon sequence replacements, than relationships determined by chemical mutation experiments. These may only be coding relations and not be directly related to viability and preservation of structure.) Since it is believed that a helical disruption must necessarily occur if a proline is present, this would mean that in a variant without proline, the helix-disrupting function must be assumed by other amino acids. When we examine proline replacements in the hemoglobins and myoglobin of Table I, we see that the -Pro-His- pair at positions 44 and 45 in alpha hemoglobin is replaced in the other molecules by -Glu-Ser-, -Asp-Ser-, and by -Asp-Asp-. In beta hemoglobin, sequence 50 through 52, -Thr-Pro-Asp-, becomes -Thr-Glu-Ala- in myoglobin. At the beginning of the E helix in beta and gamma hemoglobin there is the sequence -Asn-Pro-Lys- (positions 57 through 59) which is replaced in myoglobin by -Ser-Glu-Asp-. The proline at position 87 in the alpha chain is in that section of the sequence mentioned earlier that behaves irregularly,—the region of the heme interaction; clearly proline substitutions involving critical groups in this instance are not evident.

Finally we see that one or two prolines at the beginning of the H helix are replaced by -Ala-Asp- in myoglobin. Thus we see glutamic acid, aspartic acid, or both appearing regularly in the variant form when proline is replaced, suggesting again that they may have a similar helix-disrupting function. There are no cases in which histidine appears in place of a proline, but this may only be because of the small number of cases examined.

There are three interesting chemically induced viable mutations of tobacco mosaic virus protein that also relate to proline substitutions (Wittmann and Wittmann-Liebold, 1963). The complete sequence of this protein will not be given here, but it is found that in the section 18 through 21, Pro-20 is replaced by leucine in a nitrous acid mutant; *i.e.*, -Ala-Asp-Pro-Ile- becomes -Ala-Asp-Leu-Ile-, but in this case it is seen that Asp is present at position 19 and might suffice to disrupt a helical section if there is one. In addition, proline 63 in the sequence -Phe-Pro-Asp- can be replaced by serine in another viable mutation, but here again, Asp-64 may suffice to disrupt a helical section. It is true that proline 156 can be replaced by leucine in another mutant, and that there is no critical group nearby in the sequence. However, Fraenkel-Conrat describes this particular mutant as highly susceptible to digestion by chymotrypsin and "distinctly less viable" (Fraenkel-Conrat, 1964).

Origin of Secondary Structure. If we look for common features of the histidine, glutamic, and aspartic acid side chains that might account for their helix-disrupting behavior, they may be found in their basic and acidic character and their relatively short lengths and general lack of hydrophobic character. The maximum distances of the polar head of each of these groups from the backbone peptide chain are given in Table III, and were obtained by direct measurement from space-filling molecular models. The critical groups are seen to have shorter distances than any of

TABLE III

THE MAXIMUM EXTENSIONS OF THE CHARGED AMINO ACID SIDE CHAINS FROM THE PEPTIDE BACKBONE

Amino acid	Maximum extension of polar center from peptide backbone
	A
Asp	3.0
His	3.9
Glu	4.4
Lys	5.5
Arg	6.0

the other charged side chains. To understand why shortness and the other properties might produce a helix disruption requires some discussion of why helices form in general.

Until recent years the configuration of proteins in solution, including their helical organization, was treated independently of any solvent interactions. The helical regions were considered to be held in this specific configuration by intrachain hydrogen bonding; and these tube-like helical structures were then supposed to be held in the proper spatial configuration by salt linkages or interhelix hydrogen bonds. However, model compound studies of hydrogen bonding (Bamford *et al.*, 1956) indicate that the stabilization energy ordinarily associated with the hydrogen bond—approximately 5 kcal/mole—is greatly reduced in aqueous solution and contributes only marginal stability to the helix. The arguments against salt links have been presented by Jacobsen and Linderstrom-Lang (1949).

If we consider a polypeptide composed of predominantly non-polar amino acids, then hydrophobic bonding between the groups may take place. The free energy of stabilization of such binding is not derived from an attractive force between the bonded groups, but rather from the entropy gain that accompanies their removal from the water atmosphere (Kauzmann, 1959). In other words, water is supposed to assume a more highly ordered, hydrogen-bonded structure in the vicinity of a non-polar group. Upon removal of such hydrophobic groups from the water atmosphere, the free energy decrease then results from the entropy increase accompanying the disorganization of the water structure.

We expect then, that if hydrophobic bonding is evoked, a compact structure, although not necessarily an alpha helix, will be favored over an extended, exposed chain in aqueous solution. In order to rationalize the existence of the alpha helix, specific forces must exist, for instance, hydrogen bonding between the peptides. The assumption of hydrophobic bonding which excluded water from the peptide region, implies that the vicinity of the amide-carbonyl hydrogen bond will be partially

anhydrous and that we may expect stronger intrahelix hydrogen bonding than predicted on the basis of the model compound studies referred to above.

Our original assumption that the interaction energies involved in maintaining the secondary structure are greater than those involved in forming the tertiary structure may have a basis in the above argument. Thus to a first approximation, hydrophobic bonding that creates partially anhydrous regions plus multiple hydrogen bonding between the peptides in these regions may be responsible for the stabilization of the secondary structure; while only hydrophobic bonding conforming to the constraints due to disulfide cross-link formation stabilizes the tertiary structure.

Consider now, the introduction of a highly polar group into the peptide chain. In Fig. 1, let R_o be aspartic acid. Clearly this group exerts strong orienting forces on the water structure in its immediate vicinity and according to Frank and Wen (1957) immobilizes this layer of water. At slightly greater distances the normal water structure is altered producing a disordered region. It is possible that hydrophobic bonds between the non-polar groups have more difficulty in forming in these regions since the water structure is too strongly perturbed and no entropy gain can be realized. In addition, water cannot be excluded from the immediate neighborhood of the polar group because of the strong forces maintaining the hydrous atmosphere. Thus with no hydrophobic bonding and weak hydrogen bonding, since now water is available to compete for the carbonyl and amide groups, the helix structure may then be weak near this polar group. Longer polar side chains such as arginine and lysine may not affect the water atmosphere in the region close to the backbone chain and therefore may have less helix-weakening effect. In addition these side chains have considerable hydrophobic character themselves because of their additional methylene links occurring between the helix and the polar head.

Tertiary Structure Considerations. To return to the question of "critical groups," we see that while these may be necessary for helix disruption, it is apparent from Table I that, except for proline, they are not sufficient for this purpose. There are several positions in myoglobin and hemoglobin where histidine, glutamic acid, and aspartic acid occur in helical regions. Of course, the x-ray structure of the hemoglobins is only known to 5.5 A resolution and it is possible that the division into helical and non-helical sections does not exactly parallel the myoglobin case. In other instances when the critical groups are located in helical sections, it may also be possible that there exists a thermodynamic equilibrium distribution of secondary structures differing slightly in energy, each making use of different critical groups to divide the helical from the non-helical regions. Depending upon the energy difference between the tertiary configurations one might find a helical disruption occurring at critical group X in one of these molecular species and not in another. Presumably the form observed in x-ray crystal analysis would be the most stable; this might or might not have all the helical disruptions that are possible in solution.

Thus, in a given case, if we know what appear to be the longer structural units

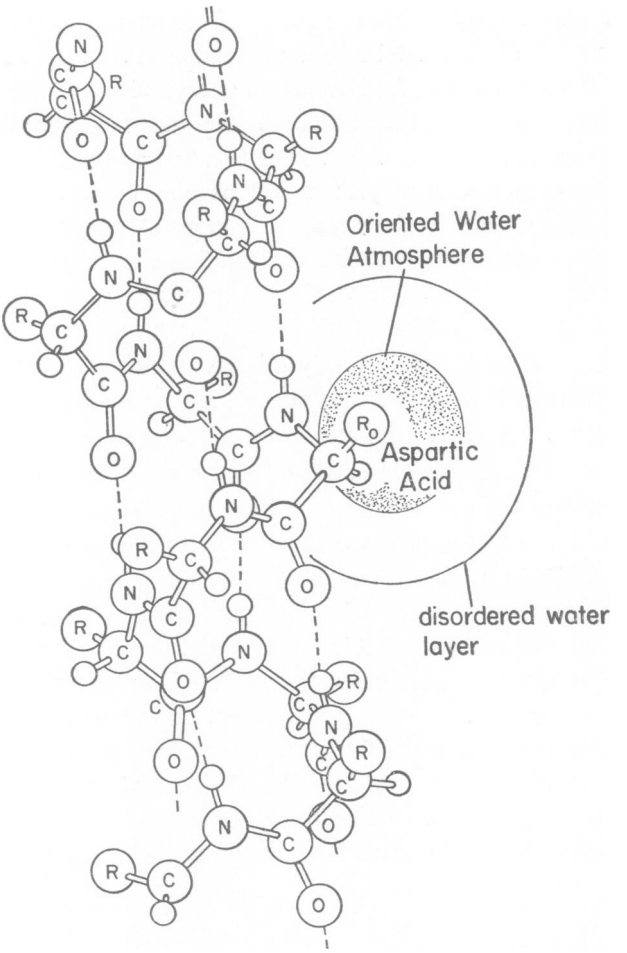


FIGURE 1 A section of an alpha helix containing an aspartic acid group. The polar nature of the side chain is thought to attract and orient the water atmosphere in the vicinity of the backbone chain.

of the protein—the lengths of the "necessary" alpha helical sections, that is, those at least six amino acids long containing no critical groups, as determined from the sequence—it may be possible to arrive at a limited set of probable or alternative tertiary structures. The realistic, possible alternatives should be guided by considering also the over-all shape of the molecule as determined by light scattering or hydrodynamic methods, and by considering the need to keep polar groups external and large non-polar groups internal to the molecule. Sections predicted to be helical after applying the above considerations may still be disrupted, of course, by the constraints imposed by disulfide links in molecules containing such bonds.

If these rules prove well founded, they may lead to the use of models such as illustrated in Fig. 2, which consist of helical sections separated (conditionally) by critical groups. These might help to reduce the difficulty of imagining what tertiary structures are likely and might lead to a smaller number of probable alternative arrangements to consider in interpreting x-ray evidence.

The present concept would suggest that a bend might occur at every or almost

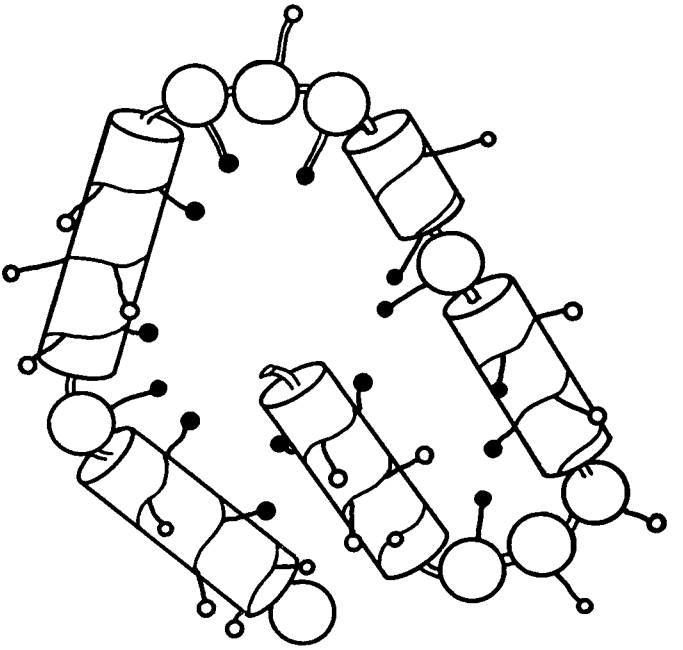


FIGURE 2 A suggested model for use in considering the relationship of helical subassemblies. The lengths of the helical sections are determined by inspection of the sequence. The arrangement should presumably be such that, in general, hydrophilic groups (indicated by the open circles) are external, and hydrophobic groups (indicated by the dark circles) are internal to the molecule.

every critical group and that sections longer than six amino acids containing no critical groups must be helical (the requirement of a minimum of six amino acids in the shortest possible helix is suggested by inspection of the shortest helices in myoglobin as given by Kendrew in Table I). This kind of suggested secondary structure might then be further refined by specific arrangements of hydrophobic groups, although it is not yet clear how to take such modifications into account.

Protein Secondary Structure. It is instructive to see if any explicit predictions can be made on the basis of this concept of critical groups. The tobacco mosaic virus (TMV) protein and lysozyme seem to be favorable cases for such a prediction.

TMV protein. The TMV protein sequence is known (Tsugita and Fraenkel-Conrat, 1963) although its secondary structure is not. The present concept would lead to the prediction that it should have the alpha helical sections given in Table IV and more schematically in Fig. 3*a*. The zigzag line represents the non-helical regions and the solid line represents the helical sections. Some of the interesting features of this construction are consistent with several peces of data that are already available. Thus we see that one very long section—helix B, 20 through 53—is predicted to be helical and contains 34 residues. If we assume the alpha helical parameters, this would imply a structure about 50 A long, a sort of longitudinal

TABLE IV

THE SUGGESTED HELICAL REGIONS OF THE TMV PROTEIN AND THE SUGGESTED AND OBSERVED HELICAL REGIONS IN LYSOZYME

Tobacco mosaic virus protein		Lysozyme		
Helix	Predicted helix sequence	Helix	Predicted helix sequence	Observed helix sequence
A	7–19	A	7–15	7–14
B	20–53	B	18–35	25–35
C	66–77	C	36–48	—
D	78–87	D	52–66	—
E	88–100	E	79–87	80–85
F	116–124	F	88–101	90–100
G	131–145	G	103–118	105–115
H	146–155	H	119–129	119–125

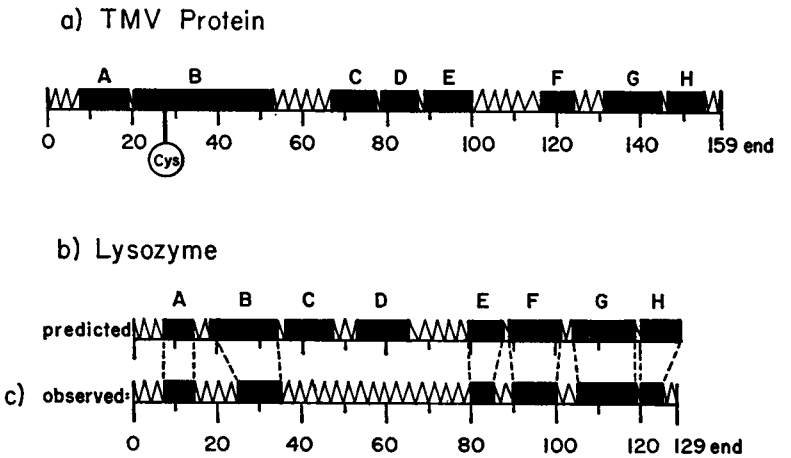


FIGURE 3 Suggested secondary structure outlines for TMV protein and lysozyme and the observed structure for lysozyme. The solid line indicates the suggested helical sections and the broken line indicates the non-helical regions.

spar or strut. It is interesting that the observed gross structure of the protein is ellipsoidal with over-all dimensions 25 × 70 A (Casper, 1963), and is thus consistent with the notion of a long helix of this kind. The cysteine group at position 27 has been located by x-ray techniques using isomorphous metallic derivatives and is found to be 16 A from one end of the ellipsoid. In the above model this group is seen to be 10 A from the end of what is predicted to be the predominant helical structure; so that the idea of an alpha helical spar composed of the region 20 through 53, running from one end to the other end of the structure, may not be an unreasonable suggestion.

Lysozyme. If we outline the sequence of lysozyme (Canfield, 1963) noting the positions of every critical group, the secondary structure given in Fig. 3*b* would be predicted, with helical sections as in Table IV.

Very recently a complete x-ray structural determination has been carried out (Blake *et al.,* 1965), and the observed secondary structure is that given in Fig. 3*c*. It is clear that except for the two helical sections C and D which are predicted but not observed, Figs. 3*b* and 3*c* are remarkably similar. Of the eight predicted helical regions, six are observed and the helical disruptions occur for the most part quite close to the positions of the critical groups. This modest success on a protein different from those in which the present critical groups were determined, suggests that this kind of subassembly partition of the structural problem deserves to be kept in mind as other protein structures are discovered.

SUMMARY

Each amino acid in the sequence of myoglobin, alpha and beta hemoglobin was classified as occurring in helical or non-helical regions of these proteins as determined from their known secondary structures. This distribution suggested a possible correlation between certain amino acids in the sequence and the location of the helical and non-helical parts of the structure, that is the presence in the sequence of one of the four "critical groups," proline, aspartic acid, glutamic acid, and histidine appears to be necessary (although the last three are not sufficient) for a helical disruption to form. It is further suggested that a length of the chain at least six amino acids long containing no "critical" groups should be helical.

These rules were applied to the sequences of tobacco mosaic virus protein and lysozyme, and secondary structures derived which are consistent with several pieces of evidence relating to the actual secondary structures of these proteins.

I would like to thank Professors Jehuda Feitelson and John R. Platt for many helpful and enlightening discussions.

This work was supported by a grant from the United States Public Health Service, No. 5 T1 GM 780-07.

Received for publication, May 17, 1965.

REFERENCES

ANFINSEN, C. B., 1961, *Proc. Nat. Acad. Sc.*, **47**, 1230.

BAMFORD, C. H., ELLIOT, A., and HANBY, W. E., 1956, Synthetic Polypeptides, New York, Academic Press, Inc.

BLAKE, C. C. F., KOENIG, D. F., MAIR, G. A., NORTH, A. C. T., PHILLIPS, D. C., and SARMA, V. R., 1965, *Nature*, **206**, 757.

BLOUT, E. R., 1962, *in* Polyamino Acids, Polypeptides, and Proteins, (M. Stahmann, editor), Madison, University of Wisconsin Press, 275.

BRAUNITZER, G., HILSE, K., RUDLOFF, V., and HILSCHMANN, N., 1964, *Adv. Protein Chem.*, **19**, 1.

CANFIELD, R., 1963, *J. Biol. Chem.*, **238**, 2699.

CASPER, D., 1963, *Adv. Protein Chem.*, **18**, 1.

DAVIES, D. R., 1964, *J. Mol. Biol.*, **9**, 605.

EDMUNDSON, A. B., 1965, *Nature*, **205**, 883.

EPSTEIN, C. J., GOLDBERGER, R. F., and ANFINSEN, C. B., 1963, *Cold Spring Harbor Symp.*, **28**, 439.

FRAENKEL-CONRAT, H., 1964, *Scient. Am.*, **211**, 46.

FRANK, H. S., and WEN, W., 1957, *Discussions Faraday Soc.*, **24**, 133.

JACOBSEN, C. F., and LINDERSTROM-LANG, K., 1949, *Nature*, **164**, 411.

KAUZMANN, W., 1959, *Adv. Protein Chem.*, **14**, 1.

KENDREW, J. C., DICKERSON, R. E., STRANDBERG, B. E., and DAVIES, D. R., 1960, *Nature*, **185**, 422.

KENDREW, J. C., 1964, private communication.

LEVINTHAL, C., 1962, *Proc. Nat. Acad. Sc.*, **48**, 1230.

NEMETHY, G., STEINBERG, I. Z., and SCHERAGE, H. A., 1963, *Biopolymers*, **1**, 43.

PERUTZ, M. F., ROSSMAN, M. G., CULLIS, A. F., MUIRHEAD, H., and NORTH, A. C. T., 1960, *Nature*, **185**, 416.

TSUGITA, A., and FRAENKEL-CONRAT, H., 1963, *in* Molecular Genetics, (J. H. Taylor, editor), New York, Academic Press, Inc., 486.

WITTMANN, H. G., and WITTMANN-LIEBOLD, B., 1963, *Cold Spring Harbor Symp.*, **28**, 589.