*PREDICTION OF α-HELICAL REGIONS IN PROTEINS OF KNOWN SEQUENCE**

By Barbara W. Low, F. M. Lovell, and Andrew D. Rudko

DEPARTMENT OF BIOCHEMISTRY, COLLEGE OF PHYSICIANS AND SURGEONS, COLUMBIA UNIVERSITY

Communicated by Erwin Chargaff, May 29, 1968

Experimental determination of detailed protein conformation by the only method now available, X-ray crystal structure analysis, depends on the preparation of suitable crystalline heavy-atom modifications if isomorphous replacement or heavy-atom methods are to be employed. The difficulties which inhibit structural determination are usually chemical problems, as encountered with proteins unresponsive to conventional methods of modification. Several new approaches, both chemical and nonchemical, to the determination of protein structure have been formulated in this laboratory and are under investigation.

The use of rigid-body search procedures[1] is one potential method under consideration. For this purpose, the conformation of some specific part of a molecule must be independently established or identified. The experimental determination of the conformation of fragment peptides, isolated regions of a protein chain, could provide appropriate search groups if it were established that regions of constant specific conformation may exist independent of environment in peptides. To investigate this possibility for insulin, an X-ray crystallographic study of fragment peptides has been begun in this laboratory. The correct prediction of conformation in certain regions of proteins would be equally valuable in providing a model for a rigid-body search study. This communication describes the development and application of criteria for the prediction of α-helical regions in proteins.

The general theory of protein structure that molecular conformation is directly dependent upon amino acid sequence has been the basis of several attempts to predict protein conformation. Thus, Scheraga and his colleagues have developed and are continuing to develop procedures to calculate the most stable conformations of single-chain proteins by employing appropriate energy functions and considering all possible intrachain and chain-solvent interactions.[2] The most common approach has been the development of criteria to predict regions of α-helix from known chain sequences. This depends on the further hypothesis that, in certain regions of a peptide chain, near-neighbor sequence interactions dominate and control local conformation. While such regions may, in principle, be either helical or nonhelical, the prediction of regions of α-helix is a conformation-defining procedure and one therefore that provides positive and potentially more useful information than does the prediction that a region is nonhelical.

Three essentially different predictive procedures have been described by other investigators. One of these was initiated by Guzzo and developed further by Prothero and by Cook.[3] This procedure depends primarily on the characterization of certain specific residues as being α-helix destabilizers on the basis of a statistical analysis of their distribution in helical or nonhelical regions of proteins of known sequence and conformation. Residues are further classified as they occur preferentially in the N-terminal and C-terminal regions of an α-helix. The predictive procedure of Schiffer and Edmundson[4] is based on the observations that intrahelical interactions involve residues n and $n \pm 3$, or $n \pm 4$,[5] that these are helix-stabilizing interactions when the residues involved are hydrophobic,[6] and that interhelical stabilizing interactions are favored by the presence of cylindrical hydrophobic arcs.[7] Appropriate allowance is made in their procedure for the apparent helix-terminating properties of certain residues. A third set of predictive criteria which depend on the calculation of the helical potential of each specific amino acid residue in its particular peptide sequence environment has been described by Periti, Quagliarotti, and Liquori.[8]

The predictive criteria to be described here differ from those described earlier, although they are also based experimentally on a search for correlations between protein conformations established by X-ray crystal structure analysis and primary amino acid sequence information. The theoretical basis of the study is the simple observation that if helix-forming sequences in which local interactions predominate should exist and can be recognized, then the particular protein in which they occur and their position along the peptide chain of that protein is by definition irrelevant.

We are essentially seeking to define the helical potential of specific sequences common to proteins of known and unknown conformation. The first stage in this study was therefore a systematic search for sequence identities within and between all the peptide chain sequences[9] of proteins of established conformation (myoglobin, lysozyme, ribonuclease A, ribonuclease S, chymotrypsin,[10] horse hemoglobin, and lamprey hemoglobin[11]) and those of a few other proteins of unknown conformation now being studied by X-ray crystallographic methods.

A simple search procedure was devised and a program written that compared systematically all the sequences in all the peptide chains of the proteins studied and pointed out coincidences. Many sequence identities were found with a broad distribution between proteins of known and unknown conformations. The search, described below, thus provided adequate experimental data for study. The analysis of these data was encouraging and a procedure for the prediction of α-helical regions was developed.

Search Procedure.—A Fortran program was written to search through the N protein chains studied for sequence identities of lengths varying between di- and hexapeptides. Let there be n_i residues in the ith chain, where $i = 1, N$. A search sequence chosen from this ith chain is then defined as the sequence of $(m + 1)$ consecutive residues,

$$S_{j,\, m}(i) = \{r_j(i), \ldots, r_{j + m}(i)\} = S,$$

where $r_j(i)$ is the jth residue of the ith chain and $j = 1, n_i$. The search is made by comparing the sequence S with the test sequences,

$$T_{k,\, m}(i) = \{r_k(i), \ldots, r_{k + m}(i)\},$$

for $k = 1, n_i$ and $i = 1, N$.

The initial value of m chosen as $m = 5$ (hexapeptide sequence) proved an appropriate upper limit. When a coincidence was found, the search sequence was identified and printed out and the sets of parameters (k,i) defining coincident test sequences were tabulated.

Results: The search yielded many identities, including one α- and β-hemoglobin chain octapeptide sequence in structural register.[11] Otherwise, the longest were three pentapeptide sequences. Two of these were between horse hemoglobin α-chain (1–5; 98–102) and lamprey hemoglobin (10–14; 114–118) and correspond, according to the Perutz sequence register, to structural homology or near-structural homology.

The third and most interesting of the three identities is the pentapeptide sequence Ala.Ala.Lys.Phe.Glu found in ribonuclease and lysozyme. It forms a part (5–9) of the first helical region in both ribonuclease A (5–12) and ribonuclease S (2–12). In lysozyme (31–35), the first four residues are in the C-terminal sequence of an α-helical region (24–34).

There were 14 tetrapeptide identities, apart from several structurally homologous pairs in the globins, and the six sets derived from the three pentapeptides. In ten pairs where conformations of both proteins are known or are proposed by sequence register, five pairs have matched conformation, all α-helical or all nonhelical. In one further pair, the lysozyme 3_{10}-helix sequence matches an α-helical sequence. One pair has a terminal residue conformation discrepancy. In only three pairs did the conformation differ completely. That seven pairs out of ten should match is persuasive.

A simple enumeration of the tri- and dipeptide identities found would take too much space. They are numerous and broadly distributed among the various proteins. Specific tripeptide and dipeptide sequences often occur more frequently than in pairs. The most frequent tripeptide sequence (excluding structural homologues) was Leu.Leu.Ser, which occurs five times in four different proteins. It is helical four times in three proteins, and nonhelical once in chymotrypsin. The most frequent dipeptide sequence (excluding structural homologues) is Ala.Ala, which occurs 20 times in seven different proteins: 13 times in helical regions of six different proteins; 4 times in nonhelical regions of two proteins, and 3 times in papain, a protein of unknown conformation.

The examples discussed are not exceptional. A detailed comparative search of all the sequence identities found strongly supported the hypothesis that sequences may be described in terms of their helical potential. As might be expected, the dominant character of a specific sequence appears more evident the longer it is. The study shows that there is considerable merit to the view that specific sequences may, in some regions of chain, determine the local conformation, at least whether it be helical or nonhelical.

Predictive Criteria and Procedures.—The study of a protein chain was made by plotting, on a map of its chain sequence, all observed identities with proteins of known conformation. The particular symbols employed showed the source protein and indicated whether the sequence was helical or nonhelical in that protein. A modified excerpt is shown in Figure 1. From inspection of the first maps prepared, it was evident that some regions of chains were mapped by overlapping sequences from the atlas of identities.

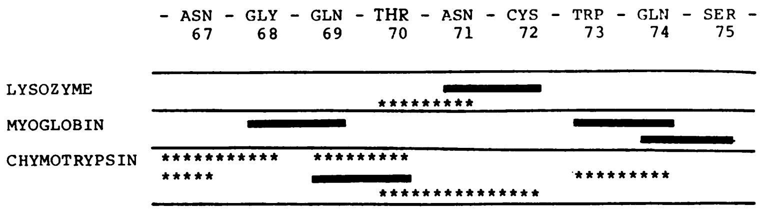

FIG. 1.—An excerpt from the ribonuclease sequence plot. (▬) Helical regions; (***) nonhelical regions.

A set of *ad hoc* criteria was developed. It was predicted that a tetrapeptide sequence or longer chain region would be helical if it were mapped by overlapping "helical" peptide sequences, e.g., for the sequence A–E to be helical, AB, BC, CD, and DE should all be found in helical conformation. The occurrence of these same sequence identities and/or other local sequence identities in nonhelical regions of proteins was completely ignored. A single occurrence of any one sequence in a helix overrode any and all occurrences in nonhelices, whether in the same or different proteins. There were modifying auxiliary rules: (1) Intrachain and interchain half-cystine residues constitute a discontinuity as they introduce an evident and unpredictable long-range effect. (2) Prolines are to be excluded from stereochemically inappropriate positions in the helix (Low and Edsall).[5] (3) A discontinuity will be bridged if the single dipeptide involved is not a common nonhelical sequence and if the adjacent regions are tripeptide or longer. (4) Adjacent regions with a single residue break will be reported if the dipeptides involved occur infrequently and if the adjacent regions are tripeptide or longer.

The helical regions employed in the study may be derived, with one exception, by shortening by one residue at each end the helical regions cited in the *Observed* lines of Table 1. This was a deliberate although arbitrary allowance for termination factors. The helical regions finally employed for lysozyme are those identified in the most recently cited conservative estimate;[10] the 3_{10}-helix was designated as helical. For ribonuclease a minimal set (5–11; 27–32; 52–57) of helical sequences was derived from the two independent studies of ribonuclease A and ribonuclease S. Information concerning α-helical regions in the hemoglobins was not included in the reservoir of information, both because of the difficulties in defining precise termination points of the helices and because in practice their trial use provided a great deal of erroneous information about other proteins.

It was recognized, as they were formulated, that these predictive procedures exploit the information available in a simple, direct, and unweighted manner.

Results: As Table 1 shows, there are, in general, rather few errors but many omissions in the predictions for proteins of known conformation as compared with observations. In lysozyme, notably, the observed 3_{10}-helix is predicted to be α-helical because it is a sequence found in a ribonuclease α-helix (53–56). The lysozyme omissions (regions which are α-helical in the structure and not so predicted) are heavily mapped but predominantly by "nonhelical" dipeptides. The

TABLE 1. *α-Helical regions predicted in proteins of known sequence and conformation.*

Hen egg-white lysozyme		
	Observed:	5–15; 24–34; 80–85; 88–96; 119–122 (3_{10})
	Predicted:	7–13; 31–35 119–122
Ribonuclease A and ribonuclease S		
	Observed:	5–12; 28–35; 51–58 (A) 2–12; 26–33; 50–58 (S)
	Predicted:	1,2–7,8;* 30–31, 32–33; (50,51–53|54–56);* 121–124
Sperm whale myoglobin		
	Observed:	3–18; 20–35; 36–42; 51–57; 58–77; 86–94; 100–118; 124–148
	Predicted:	70–73
Tosyl-α-chymotrypsin		
A chain	Observed:	None
	Predicted:	None
B chain:	Observed:	None
	Predicted:	(50–52|53–56); 67–71; 85–88; 104–107; 110–113
C chain	Observed:	238–245
	Predicted:	157–159,160; 176–180; 183–186; (231–239:241–244)*

* A comma separating two sequences indicates that these would have formed one continuous array if the terminal restrictions on helical sequences employed for prediction had not been imposed. The vertical bar is used to indicate regions of predicted helix which are adjacent, but for which there is no connecting dipeptide (Auxiliary Rule 3). The colon is used to link regions of predicted α-helix, tripeptides or longer, separated by a single residue, and for which the two connecting dipeptides occur infrequently.

limited ribonuclease omissions, on the other hand, are from unpredictable regions very lightly mapped by coincident peptides from the information reservoir. The many myoglobin omissions are either lightly mapped or largely mapped with nonhelical coincidences. In chymotrypsin there are several errors. The longest sequence predicted, however, does include as terminal segment the only observed α-helical region (238–245).

When our predictions and those of other investigators are compared (Table 2), the conservative nature of this procedure and its relative freedom from errors are re-emphasized. Thus, in lysozyme, other investigators have correctly predicted more α-helical regions, but their predictions have been accompanied by very large errors. In the other proteins, our method leads always to many fewer wrong assignments. In myoglobin, where our procedures have provided so little information, the Schiffer and Edmundson criteria, based as they are on an analy-

TABLE 2. *Comparison of predictions.*

Protein	Number of α-helical residues		Predictions				
			1	2(*a*)	2(*b*)	3	4
Lysozyme	41	Correct	15	39	30		34
		Wrong	1	66	9	—	27
Ribonuclease A	24	Correct	18		19	17	15
		Wrong	5	—	19	15	18
Myoglobin	119	Correct	4		100		82
		Wrong	0	—	16	—	2
Chymotrypsin	8	Correct	7		2		4
		Wrong	44	—	115	—	91

1, This method; 2(*a*), Guzzo; 2(*b*), Prothero; 3, Periti; 4, Schiffer and Edmundson sis of the known myoglobin helical regions, provide good results. Prothero's procedures, however, lead to an even larger number of correct identifications.

Predictions for proteins of unknown conformation (Table 3) show some features common to other estimates. Thus, for cytochrome *c*, Schiffer and Edmundson[4] predict five α-helical regions and two of these, 80–87 and 93–101, overlap closely with our predictions. In papain two (77–86; 99–112) of the three regions of high-helix probability predicted by Jansonius,[12] using the criteria of Periti, Quagliarotti, and Liquori,[8] show overlap with our predictions. In mapping the papain sequences with sequences from other proteins, the region 190–198 appeared so impressively and overwhelmingly made up of "nonhelical" dipeptide sequence overlaps that we predict that this region will be nonhelical, even though we have not established general criteria for nonhelical sequences.

TABLE 3. *α-Helical regions predicted in proteins of known sequence but unknown conformation.*

Glucagon: None

Cytochrome *c:* 83–84,85–88; (92–96|97–101)

Papain:* 11–14; 37–40; 65–70; 73–76; 82–85;
(101–104:106,107–111); 118–121

Bovine insulin: A chain, none; B chain, (11–12,13–15|16–18)

Bonito insulin: A chain, (12–14|15–17); B chain, (11–12,13–15|16–18)

* The predictions tabulated here for papain are based on three independent chain sequences (1–26, 31–160, and 171–198). Uncertainties in residue assignment make prediction for the intermediate sequences impossible. The particular α-helical regions predicted differ from an earlier set presented at the *Working Conference on X-Ray Crystallography of Proteins* (Arden House, November 1967). The regions 37–40 and 118–121 were then erroneously omitted and the region 82–85 was extended to 80–87. At that time the rules permitted discontinuities to be bridged if one of the adjacent peptides was a dipeptide. The symbols used are the same as in Table 1.

If specific sequences may lead to conformation prediction, then species differences between proteins should not provide conflicting results. We therefore examined the chain sequences of bovine, sheep, horse, sei whale, human, bonito, elephant, rabbit, rat (2), chicken, guinea pig, toad fish (2), and angler fish (B chain only) insulins. None of the insulin sequences except bonito leads to helix prediction in the A chain, although three other insulins differ from bovine insulin in this region. The prediction of B chain 11–18 as the sole α-helical sequence for bovine insulin is maintained for all the insulins except guinea pig, where the region 11–18 is broken by a two-residue gap.[13] Schiffer and Edmundson[4] predict three regions of α-helix in bovine insulin, A1–6, 12–20, and B9–19.

Comments.—The aim of this study to provide reliable and conservative predictions of α-helical regions has been in part realized. Simultaneously, its more proper purpose to predict helical potential has been recognized. Because the procedure adopted does lead to conservative estimates of α-helical structure, evidently predictions of high helical potential and of α-helix do frequently coincide. The basic and novel assumption that the helical potential of a sequence can be derived by considering the helical potential of its component shorter sequences appears, generally, to be validated; the anomalous prediction of α-helix for the lysozyme 3_{10}-loop is itself paradoxically forceful evidence of this.

The proteins now studied and employed are few; it cannot therefore yet be shown that the observations are generally valid and the method is therefore generally applicable. If both the protein sequences studied and the sequence-conformation relationships in the information pool are random representative samples of all proteins, then the power and range of the method will increase as more conformations are established.

Reliable predictions may be of value in several contexts. They can provide: (*a*) time-saving information useful in general calculations of most stable chain conformation; (*b*) guide lines for identifying both main-chain direction and specific residues in low-resolution electron density maps of proteins, and (*c*) rigid-body search groups in vector structure analyses. Although the minimum length of α-helix appropriate for use with such procedures cannot be generally defined, a nona- or decapeptide length might merit investigation with a small protein.[14]

To improve the procedure, "errors" must be reduced and omissions repaired. Whatever the helical potential of a tetrapeptide sequence, it cannot maintain a single turn independent of interactions with other regions adjacent or nonadjacent to it. As Schellman and Schellman[15] have pointed out, this is true of even longer sequences, as only $N - 8$ residues of an N-residue length α-helix are genuine helical residues with both NH and CO held by intrahelical hydrogen bonds. Short regions of α-helix are found in proteins. Indeed, the average α-helix in myoglobin, lysozyme, ribonuclease, and chymotrypsin is only 11 residues long, and without myoglobin the average is only 8.

Modification of the predictive procedure should:

(1) Recognize and allow for the degree of helical potential of a sequence, rather than define its absolute helical or nonhelical character. A weighting procedure should take account of the over-all frequency of occurrence of a sequence and its position in (*a*) an α-helix at the N- or C-terminal ends or in the true helix "core," and (*b*) in a nonhelical conformation.

(2) Provide estimates of the helix-stabilizing and destabilizing effects of residues near-neighbor to regions of high helical potential and thus permit lengthening or elimination of short predicted regions.

Development of weighting schemes could make possible parallel predictions of nonhelical conformation as were made for papain. If reasonably long nonhelical sequences common to two proteins are found, then the related question of specificity of nonhelical conformation can be explored. Although the β structure is by definition helical, it appears improbable that the method will be valid or extensible to this structure which is characterized in terms of non-near-neighbor interactions.

We have shown that local sequence character is important. To the extent that helical stability is wholly dependent on near-neighbor or adjacent segment character, conformation is determinable and the predictive criteria theoretically perfectable; only where helical regions are critically dependent for their stability on interactions between segments of chain far removed in sequence are our criteria inappropriate.

* This investigation was supported by U.S. Public Health Service research grant RO1-AM-01320 from the National Institute of Arthritis and Metabolic Diseases.

[1] Nordman, C. E., and K. Nakatsu, *J. Am. Chem. Soc.*, **85**, 353 (1963); Nordman, C. E, *Trans. Amer. Crystallogr. Assoc.*, **2**, 29 (1966).

[2] Scheraga, H. A., in *Advances in Physical Organic Chemistry*, ed. Victor Gold (London: Academic Press, 1968).

[3] Guzzo, A. V., *Biophys. J.*, **5**, 809 (1965); Prothero, J. W., *Biophys. J.*, **6**, 367 (1966); Cook, D. A., *J. Mol. Biol.*, **29**, 167 (1967).

[4] Schiffer, M., and A. B. Edmundson, *Biophys. J.*, **7**, 121 (1967).

[5] The importance of interactions in the α-helix between residues n and $n \pm 3$, or $n \pm 4$, has been remarked; see Low, B. W., and J. T. Edsall, in *Currents in Biochemical Research 1956*, ed. D. E. Green (New York: Interscience Publishers, Inc., 1956), pp. 378–433; Low, B. W. in *Molecular Biology*, ed. D. Nachmansohn (New York: Academic Press, 1960).

[6] Némethy, G., and H. A. Scheraga, *J. Phys. Chem.*, **66**, 1773 (1962).

[7] Analysis of the globin chain structure and sequence has shown that in helical regions nonpolar sites tend to repeat at regular intervals of about 3.6 residues, thus forming cylindrical nonpolar arcs; see Perutz, M. F., J. C. Kendrew, and H. C. Watson, *J. Mol. Biol.*, **13**, 669 (1965).

[8] Periti, P. F., G. Quagliarotti, and A. M. Liquori, *J. Mol. Biol.*, **24**, 313 (1967): Periti, P. F., *Nature*, **215**, 509 (1967). A statistical analysis of the sequence/conformation relationships in proteins of established tertiary structure has been made and the results tabulated. For the proteins studied, weighting factors are calculated for each residue based on the number of times it and the specific residue which precedes or follows it by six or fewer places occur in like relative positions to each other as parts of helical or of nonhelical arrays.

[9] The primary sequences employed were taken from Eck, R. V., and M. O. Dayhoff, in *Atlas of Protein Sequence and Structure 1966* (National Biomedical Research Foundation, 1966). All were checked against the original publication cited there.

[10] Myoglobin: Kendrew, J. C., R. E. Dickerson, B. E. Strandberg, R. G. Hart, and D. R. Davies, *Nature*, **185**, 422 (1960); lysozyme: Blake, C. C. F., G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, *Proc. Roy. Soc.* (*London*), **B167**, 365 (1967); ribonuclease A: Kartha, G., J. Bello, and D. Harker, *Nature*, **213**, 862 (1967); ribonuclease S: Wyckoff, H. W., K. D. Hardman, N. M. Allewell, T. Inagami, D. Tsernoglou, L. N. Johnson, and F. M. Richards, *J. Biol. Chem.*, **242**, 3749 (1967); tosyl-α-chymotrypsin: Matthews, B. W., P. B. Sigler, R. Henderson, and D. B. Blow, *Nature*, **214**, 652 (1967).

[11] The helical regions in the α and β chains of horse hemoglobin and those in the chain of lamprey hemoglobin have been "established" by analogy with myoglobin from the sequence register and low-resolution Fourier; see Perutz, M. F., *J. Mol. Biol.*, **13**, 646 (1965), and Cullis, A. F., H. Muirhead, M. F. Perutz, M. G. Rossmann, and A. C. T. North, *Proc. Roy. Soc.* (*London*), **A265**, 161 (1962).

[12] Jansonius, J. N., in *De Kristalstructuur van Papaïne* (Groningen: Rijksuniversiteit Te Groningen, 1967), p. 121.

[13] In the toad fish and the angler fish the extra residue at the N-terminal of the B chain is considered, following Smith's convention, as −1B; this maintains the over-all sequence register (Smith, L. H., *Am. J. Med.*, **40**, 662 (1966)).

[14] Using his rigid-body search method, Nordman has found the orientations of the five longest α-helices of myoglobin in the vector structure of the crystal (private communication). The regions are 25, 20, 19, 16, and 16 residues long.

[15] Schellman, J. A., and C. Schellman, in *The Proteins*, ed. H. Neurath (New York: Academic Press, 1964), vol. 2, p. 1.