LETTERS TO THE EDITOR

Correlation between the Distribution of Amino Acids and Alpha Helices

Dear Sir:

In a recent paper in this *Journal* (1), Guzzo proposed, implicitly if not explicitly, a tentative rule for predicting the distribution of α-helices from amino acid sequences. In that paper the rule was employed to predict the helical regions in tobacco mosaic virus and lysozyme. In the interest of examining rules of this general kind, it is useful to restate his rule perhaps more explicitly than originally intended and to apply it to other proteins. The rule states in effect that an α-helical region of a polypeptide chain may include a single proline, aspartate, glutamate, or histidine residue only at each end (but proline only at the α-amino end) of a sequence containing at least six residues other than proline, aspartate, glutamate, or histidine. Thus the latter four residues are said to function as α-helix destabilizers. Proline is known to disrupt α-helices whereas it is assumed that aspartate, glutamate, and histidine are only conditional destabilizers, the conditions remaining to be specified. One evident merit of this rule is its simplicity. Furthermore, if the rule has real predictive value, it will be of considerable theoretical interest inasmuch as it would possibly be an important step toward calculating the tertiary structure of a protein directly from its amino acid sequence. But the central question, in view of the limited evidence available, is whether this rule (or any other given rule) does, in fact, have predictive value.

Strictly, only the published data for sperm whale myoglobin (2) and lysozyme (3) enable a direct test of a rule to be made at the present time. Nevertheless, in view of the close similarities between hemoglobin and myoglobin, it is reasonable to employ the available hemoglobin data (4) for this purpose as well. These examples represent only a small sample of all proteins. It is both desirable and possible to test a given rule against those other proteins whose amino acid sequences are known and whose α-helical content has been estimated by optical rotatory dispersion. There is good agreement between the α-helical content observed by optical rotatory dispersion measurements and that determined by x-ray structure analysis for both myoglobin and lysozyme. It now seems unlikely that the optical rotatory dispersion measurements give a grossly wrong result. Amino acid sequence data and estimates of helical content are available at least for pancreatic ribonuclease (5, 6), tobacco mosaic virus (5, 6) and bovine chymotrypsinogen A (6, 7). A simple procedure then to test a rule is to first calculate the helical segments and then the predicted total helical content. This may be compared with the observed value. Evidently it is desirable to carry out this calculation for myoglobin, lysozyme, and hemoglobin as well. Note that a rule might predict 42% helix for lysozyme but, nevertheless, be quite wrong! It is essential to compute wherever possible some objective measure of the "goodness of fit" achieved by a given rule. A simple measure, suitable for proteins having at least a moderate α-helical content is given by:

$$\text{Goodness of fit} = \frac{\text{No. right—No. wrong}}{\text{total no. helical residues}} \times 100$$

where the "No. right" is the number of α-helical residues correctly predicted. Accordingly, 100% corresponds to a rule achieving a perfect fit and 0% to a rule that is wrong as often as it is right. Where more are wrong than right the index will be negative. At present this measure may be used to test predictions about myoglobin, lysozyme, and hemoglobin. The helical segments, per cent helix, and goodness of fit have been calculated according to the explicit statement of the rule cited above. The results are summarized in Table I. With the exception of two short segments (see footnote to Table I),

TABLE I

THE RULE (1) CITED IN TEXT FOR PREDICTING α-HELICAL DISTRIBUTION FROM AMINO ACID SEQUENCE MAY BE PARAPHRASED "THAT WHENEVER SIX OF US ARE GATHERED TOGETHER WE SHALL BE α-HELICAL."

"We" does not include proline, aspartate, glutamate, or histidine. The rule has been used to calculate the α-helical segments, per cent α-helix and goodness of fit (i.e. No. right minus No. wrong over total) for myglobin, lysozyme, the α-, β-, and γ-hemoglobin chains, ribonuclease, chymotrypsinogen A, and tobacco mosaic virus protein.

Protein	Predicted helical segments	Helix		Goodness of fit
		Predicted	Observed	
		%	%	%
Mb	27–36, 64–81, 126–136, 141–148	31	77	33
Lysozyme	7–15, 18–35, 36–48, 52–66, 70–78*, 79–87, 88–101, 103–119, 120–129	88	39	−36
Hb-α	9–20, 30–36, 37–44, 78–85, 103–112, 126–141	43	74	47
Hb-β	7–19, 26–36, 63–73, 80–90, 108–116, 131–139	44	77	46
Hb-γ	7–19, 26–36, 47–57, 63–73, 80–87, 108–117, 131–146	55	77	44
RNase	2–9, 14–38, 53–83, 86–92, 93–105, 106–113	74	17	—
Chymotrypsinogen A	8–19, 28–35, 39–48, 49–56, 57–63, 64–70, 78–123, 129–151, 153–160, 161–178, 179–194, 198–225, 226–246	86	11	—
TMV	7–19, 20–53, 56–62*, 66–77, 78–88, 89–97, 116–125, 131–145, 146–155	76	30	—

* These helical segments appear to be predicted by Guzzo's rule but are not given in his paper [see (1), Table IV].

the predicted helical segments for tobacco mosaic virus and lysozyme are essentially the same as those given (1). It should be noted that the values of per cent helix given in Table I are to be regarded as minimum values inasmuch as aspartate, glutamate, and histidine are considered in the rule cited above as conditional destabilizers.

TABLE II

Protein		Helical segments								α-helix		Goodness of fit
										Observed	Predicted	
										%	%	%
Human Hb α-chains	observed	5–18	22–36	37–42	51–52	54–72	81–88	95–112	119–141	74	67	68
	predicted	7–14	17–34		51–57	60–75	77–88	100–113	118–126			
									128–138			
Human Hb β-chains	observed	6–19	21–35	36–41	51–57	59–78	86–94	100–117	124–146	77	60	69
	predicted	4–19	20–35			66–71	84–93	104–117	124–144			
						74–81						
Myoglobin	observed	3–18	20–35	36–42	51–57	58–77	86–94	100–119	125–148	78	63	66
	predicted	3–18	19–24	36–44	47–61	63–78	88–95	103–116	124–140			
						80–86			142–149			
Lysozyme	observed	4–15	24–36		81–85	88–95	99–100	108–115	119–125	42	30	65
	predicted	4–14	28–35		80–86	89–95		106–111				
Human Hb γ-chains	observed	6–19	21–35	36–42	51–57	59–78	86–93	100–117	124–146	77	62	65
	predicted	3–11	17–35		48–54	65–71	82–93	104–119	123–143			
RNase	predicted	2–11			43–61			105–113		17	30	—
TMV	predicted			39–46		72–76	78–79	94–100	102–136	30	42	—
Chymotrypsinogen A	predicted	9–13			43–73	102–123	160–164	204–215		11	47	—
	predicted	15–23				82–89	131–137	179–185	228–239			

It may be remarked that in the case of a rule depending heavily upon a small number of residues, say three or four, it is worth examining those cases where only the amino acid composition (but not the sequence) and the per cent α-helix is known. An excellent summary of the data is given by Davies (6). A case in point is paramyosin, which contains 20.7 mole % glutamate, 14.0% aspartate, and 0.4% histidine, or a total of 35.1% of α-helix destabilizers according to the explicit statement of the rule cited above. But optical rotatory dispersion measurements suggest paramyosin is about 95% helical.

In closing I would like to suggest the following rule which does appear to achieve a reasonable degree of fit with the known protein structures. The rule is; any region of five residues will be α-helical if at least three of those residues comprise alanine, valine, leucine, or glutamate. Alternatively, any region of seven residues will be α-helical if at least three residues comprise alanine, valine, leucine, or glutamate and an additional one includes glutamine, isoleucine, or theronine. Proline is assumed to be restricted to the last or last but one position at the α-amino end of a helix. The three residues at each end of a polypeptide chain are ignored in calculating the helical regions. The results obtained with this rule are summarized in Table II.

It will be clear that the above rule is based in part upon the assumption that the amino acid sequences of proteins are not highly ordered. This appears to be a reasonable assumption for globular proteins, if not for fibrous proteins. Thus it is not expected that the rule will necessarily work for highly ordered polypeptides. The plausible assumption which has usually been made, namely that if a given residue [e.g. valine (8)] is a destabilizer in an order polypeptide, it will also be a destabilizer in a relatively unordered polypeptide may not be warranted. The data for the helical H-segments of hemoglobin suggests that it is a questionable assumption. Admittedly, accepting this statement is not quite the same thing as asserting that a given residue (e.g. valine) is a helix-former when it occurs in relative isolation. The reader may be able to revise the above rule to free it from this kind of criticism.

Received for publication 10 January 1966.

REFERENCES

1. GUZZO, A. V., *Biophysic. J.*, 1965, **5**, 809.
2. EDMUNDSON, A. B., *Nature*, 1965, **205**, 883.
3. BLAKE, C. C. F., KOENIG, D. F., MAIR, G. A., NORTH, A. C. T., PHILLIPS, D. C., and SARMA, V. R., *Nature*, 1965, **206**, 757.
4. BRAUNITZER, G., HILSE, K., RUDLOFF, V., and HILSCHMANN, N., *Advances Protein Chem.*, 1964, **19**, 1.
5. CANFIELD, R. E., and ANFINSEN, C. B., *in* The Proteins I, (H. Neurath, editor), New York, Academic Press, Inc., 1963.
6. DAVIES, D. R., *J. Mol. Biol.*, 1964, **9**, 605.
7. HARTLEY, B. S., *Nature*, 1964, **201**, 1285.
8. FRASER, R. D. B., MACRAE, T. P., and STEWART, F. H. C., *J. Mol. Biol.*, 1965, **13**, 949.

JOHN W. PROTHERO
Department of Biological Structure
School of Medicine
University of Washington
Seattle, Washington