Proc. Nat. Acad. Sci. USA
Vol. 70, No. 10, pp. 2809–2813, October 1973

Prediction of the Amount of Secondary Structure in a Globular Protein from Its Aminoacid Composition

(helix/β-sheet/turns)

W. R. KRIGBAUM AND SARA PARKEY KNUTTON

Gross Chemical Laboratory, Duke University, Durham, North Carolina 27706

Communicated by Walter Gordy, June 22, 1973

ABSTRACT **Multiple regression is used to obtain relationships for predicting the amount of secondary structure in a protein molecule from a knowledge of its aminoacid composition. We tested these relations using 18 proteins of known structure, but omitting the protein to be predicted. Independent predictions were made for the two subchains of hemoglobin and insulin. The average errors for these 20 chains or subchains are: helix ± 7.1%, β-sheet ± 6.9%, turn ± 4.2%, and coil ± 5.7%. A second set of relations yielding somewhat inferior predictions is given for the case in which Asp and Asn, and Glu and Gln, are not differentiated. Predictions are also listed for 15 proteins for which the aminoacid sequence or tertiary structure is unknown.**

Protein crystallographers often undertake the crystal-structure study of a protein when its aminoacid composition, but not the residue sequence, is known. Hence, a method for predicting the amounts of the various features of secondary structure (helix, β-sheet, etc.) from the aminoacid composition could be of considerable use. A second possible area of application arises when a knowledge of the sequence is used to predict the regions of secondary structure. For example, prediction of helical regions in proteins often involves, in part, comparison of the helical potential of each residue with some threshold value. The threshold is usually the same for all proteins predicted, although its value can be adjusted arbitrarily to obtain the best possible agreement. If the helical content can be estimated independently from the aminoacid composition, the threshold may be correspondingly adjusted for each protein, thereby improving the prediction. In our experience, the percentage of residues correctly predicted as being helical or nonhelical, based on single residue potentials, can be increased by more than 10% in this manner. The same procedure can, of course, be applied to prediction of regions of other types of secondary structure.

PROCEDURE

18 Proteins of known sequence and structure were the data base for our study. Independent predictions will be given for the A and B subchains of hemoglobin and insulin, bringing the total to 20. Four types of secondary structure are considered: helix, β-sheet, turn, and coil (or, more precisely, the remainder). The assignments are largely based upon the shapes of the secondary structural regions as portrayed by Rubin–Richardson wire models (1). Both α and 3_{10} helices are included in the helical category, and the criteria of Venkatachalam (2) and Matthews (3) were used in assigning turns. Flexibility of the wire models made assignment of β-sheet regions difficult, so reference was made to the published crystal structures where ambiguities arose. Regions not included in the above categories were termed coil. Assignments of secondary structural regions appear in Table 1. Since residues may be assigned to more than one category (e.g., the last residues of a helical region and the first ones of a turn), then percentages do not necessarily add up to 100% for any protein.

We recognize at the outset that this data base contains errors of at least two types. First, there are some remaining uncertainties in the primary structures. Occasionally the chemical and crystallographic identification of a particular residue differs. More frequently, the acids Asp and Glu, and their corresponding amides Asn and Gln, are not differentiated. These uncertainties adversely affect the quality of prediction. Secondly, assignment of the regions involves some measure of uncertainty. For example, we may compare the helical regions for eight common proteins listed in four of the recent papers (4–7) concerned with prediction of helical regions. These proteins were comprised of 1586 residues, and there were differences in 94 of the approximately 570 residues assigned as helical, which corresponds to an uncertainty of 6% in the helical content. For the other types of secondary structure we cannot obtain even this crude estimate, but it is probably reasonable to assume that the uncertainty will be as large or larger, since the other regions are not as well defined.

We now turn to methods for prediction. One might expect to find a positive correlation between the helical content and a parameter obtained by summing the product of the percent composition of each amino acid and the helical potential of the corresponding residue. Here the helical potential is taken as the fraction of all occurrences of that type of residue in the data set that are within helical regions. The composite parameter might be considered to represent, in effect, the "helical potential" of the entire molecule. In fact, however, the correlation between helical content and this composite parameter is very weak, and even appears to be negative. Perhaps this finding should not be surprising, since prediction of helical regions from single residue potentials also uses additional information contained in the sequence of residues.

We next examined correlations between the sum of the percentage compositions of from one to five selected types of amino acid and the percentages of the various features of secondary structure. Several combinations gave high correlations. For helical content, for example, 16 combinations yielded a Pearson correlation coefficient, r, of +0.80 or larger, and 8 gave negative coefficients exceeding 0.70 in absolute magnitude ($|r|$ ranges from 0.3 to 0.4 for random). Two sums

TABLE 1. *Assigned regions of secondary structure in eighteen proteins*

Protein	Helix	β-Sheet	Turn	Coil
Myoglobin	3-18,20-35,36-42,51-57, 58-77,86-95,100-118, 124-149		18-21,36-39,43-46,46-49, 78-81,82-85,95-98, 119-122,121-124	1,2,50,99,150-153
Oxyhemoglobin B	4-18,19-34,35-41,50-56, 57-76,86-94,99-117, 123-143		14-17,32-35,35-38,42-45, 46-49,53-56,79-82,81-84, 92-95,97-100,118-121, 120-123	1-3,77-78,85,96, 144-146
Oxyhemoglobin A	3-18,20-35,36-42,52-71, 80-89,94-112,118-138		17-20,36-39,41-44,43-46, 49-52,70-73,73-76,76-79, 89-92,114-117,137-140	1-2,47-48,93,113,141
Cytochrome b_5	8-15,33-38,42-49,55-62, 64-74,80-86	4-6,21-25,28-32,50-54, 75-79	16-19,18-21,25-28,39-42, 49-52,78-81	1-3,7,63,87
Carp myogen	8-18,26-33,40-51,58-62, 65-69,79-88,100-107		2-5,20-22,35-38,38-41, 51-54,70-73,91-94	1,6-7,19,23-25,34,55-57, 63-64,74-78,89-90,95-99, 108
Insulin B	9-20		6-9,13-16	1-5,21-30
Insulin A	2-8		6-9,13-16	1,10-12,17-21
Lysozyme	5-15,25-36,81-85,88-99, 109-114,120-124	42-61	17-20,20-22,36-39,39-42, 47-50,54-57,60-63,69-72, 74-77,85-87,103-106, 106-109,115-118,124-127	1-4,16,23-24,64-68,73, 78-80,100-102,119, 128-129
Carboxypeptidase	14-29,72-88,94-103, 115-123,174-184, 215-233,254-262, 288-305	32-37,45-54,61-67, 104-111,190-197, 200-205,239-243, 265-270	3-6,4-7,29-32,41-44, 56-59,67-70,89-92, 110-114,114-117,117-120, 123-126,142-145,148-151, 150-153,159-162,162-165, 169-172,206-209,214-216, 232-235,242-245,244-247, 273-276,277-280,283-286	1-2,8-13,38-40,55,60, 71,93,127-141,146-147, 154-158,166-168,173, 185-189,198-199,210-213, 236-238,248-253,263-264, 271-272,281-282,287, 306-307
Thermolysin	65-89,137-150,160-179, 235-246,264-273,281-295, 302-313	3-13,15-25,27-32,35-46, 52-58,60-63,97-106, 112-116,119-123	13-16,24-27,32-35,35-38, 44-47,49-52,57-60,88-91, 92-95,107-110,110-113, 126-129,127-130,132-135, 150-153,159-162,178-181, 187-190,190-193,194-197, 197-200,201-204,204-207, 205-208,209-212,217-220, 224-227,225-228,231-234, 249-252,250-253,259-262, 262-265,272-275,276-279, 294-297,298-300,300-303	1-2,48,64,96,117-118, 124-125,131,136, 154-158,182-186, 213-216,221-223, 229-230,247-248, 254-258,280,314-316
Subtilisin	5-10,14-20,64-73,103-117, 132-145,223-238,242-252, 269-275	28-32,45-50,89-94, 120-124,149-152	23-26,36-39,39-42,51-54, 56-59,60-63,61-64,83-86, 85-88,97-100,145-148, 159-162,167-170,171-174, 181-184,187-190,193-196, 210-213,219-222,238-241, 260-263,263-266	1-4,11-13,21-22,27, 33-35,43-44,55,74-82, 95-96,101-102,118-119, 125-131,153-158, 163-166,175-180, 185-186,191-192, 197-209,214-218, 253-259,267-268

of up to five amino acids (one giving a positive correlation and the other negative) were combined by a multiple regression technique. Comparison of the predictions based upon different combinations of amino acids showed that the best pair did not always involve the two sums having the largest $|r|$ values.

Preliminary examination revealed that the predicted values of the heme proteins were low relative to those for the other proteins. Our aminoacid combinations take no cognizance of the presence of the heme group. Optical rotatory dispersion measurements (8–11) indicate that apomyoglobin has 20% less helix than myoglobin, which suggests addition of 20% to the calculated helical percentages for oxygen-carrying heme proteins.

RESULTS

The combinations of aminoacid compositions found to give the best predictions are:

helix: $H^+ = \text{Ala} + \text{Leu} + \text{Glu} + \text{His} + \text{Asn}$ ($r = +0.85$)
$H^- = \text{Pro} + \text{Thr} + \text{Gln}$ ($r = -0.79$)

β-sheet: $B^+ = \text{Asp} + \text{Thr} + \text{Arg} + \text{Gln} + \text{Val}$ ($r = +0.83$)
$B^- = \text{Leu} + \text{Asn} + \text{Ala} + \text{Glu} + \text{Gly}$ ($r = -0.82$)

turn: $T^+ = \text{Gly} + \text{Thr} + \text{Asp} + \text{Glu} + \text{Asn}$ ($r = +0.75$)
$T^- = \text{Ser} + \text{Ala} + \text{Arg} + \text{Phe} + \text{Glu}$ ($r = -0.85$)

coil: $C^+ = \text{Cys} + \text{Tyr} + \text{Ala} + \text{Thr}$ ($r = +0.73$)
$C^- = \text{His} + \text{Asp} + \text{Asn}$ ($r = -0.73$)

(TABLE 1. *continued*)

Protein	Helix	β-Sheet	Turn	Coil
Staph. nuclease	54-69,99-107,121-134	12-27,30-36	1-4,4-6,19-22,26-29, 27-30,36-39,47-50,49-52, 52-55,69-72,77-80,78-81, 83-86,94-97,108-110, 116-119,120-123,133-136, 137-140,139-142	7-11,40-46,73-76,82, 87-93,98,111-115, 143-149
Ribonuclease	2-12,26-33,50-58	42-49,71-92,94-110	16-19,34-37,36-39,65-68, 67-69,87-90,92-95,112-115	1,13-15,20-25,40-41, 59-64,70,111,116-124
Cytochrome *c*	2-13,92-102		11-14,15-18,21-24,26-29, 32-35,35-38,44-46,49-51, 52-55,60-62,61-64,63-66, 67-69,68-71,71-74,73-76, 75-78,78-81,87-90	1,19-20,25,30-31,39-43, 47-48,56-59,82-86,91, 103-104
Papain	26-41,50-56,69-78,116-126	107-115,127-133,159-164, 170-174,184-189	2-4,6-9,8-11,19-22,40-43, 57-60,61-64,67-70,82-85, 84-87,90-92,97-100,98-101, 114-117,138-141,142-145, 147-150,168-171,179-181, 182-185,184-187,194-197, 196-199,199-202,209-212	1,5,12-18,23-25,44-49, 65-66,79-81,88-89, 93-96,102-106,134-137, 146,151-158,165-167, 175-178,190-193,203-208
Trypsin inhibitor	47-56	9-13,17-20,35-36,39-45	4-7,24-27,25-28,42-44	1-3,8,14-16,21-23, 29-34,37-38,46,57-58
α-Chymotrypsin	165-170,234-245	29-35,39-46,50-55, 86-91,103-108,134-141, 155-164,179-184,197-203, 206-214,226-230	23-26,27-30,48-51,55-58, 56-59,61-64,72-75,91-94, 95-98,115-118,125-128, 131-134,167-170,171-174, 174-177,177-180,185-188, 191-194,194-197,203-206, 217-220,221-224,230-233, 231-234	1-22,36-38,47,60,65-71, 76-85,99-102,109-114, 119-124,129-130,142-154, 189-190,215-216,225
Elastase	155-160,232-240	16-22,25-35,38-44, 53-58,69-80,93-101, 124-135,139-153, 171-180,185-196, 199-210,215-226	3-5,8-11,10-13,21-24, 28-31,36-39,45-48,61-64, 80-83,84-87,99-102, 100-103,106-109,116-119, 122-125,135-138,136-139, 160-163,164-167,170-172, 178-181,187-190,195-198, 213-216,217-219	1-2,6-7,14-15,49-52, 59-60,65-68,88-92, 104-105,110-115,120-121, 154,168-169,182-184, 211-212,227-231
Rubredoxin		3-7,10-20,23-30, 37-39,42-45,50-52	6-9,7-10,14-17,20-23, 25-28,29-32,34-37,36-39, 39-42,47-49,52-54	1-2,33,46
Concanavalin A		5-15,19-27,35-40, 47-55,59-66,70-80, 90-98,102-115,122-130, 146-149,152-158,163-164, 169-174,177-182,188-196, 207-215	15-18,28-31,43-46,55-58, 56-59,67-70,81-84,83-86, 86-89,98-101,115-118, 118-121,133-136,136-139, 141-144,145-147,148-151, 159-162,165-168,174-177, 182-185,202-205,215-218, 221-224,225-228	1-4,32-34,41-42,131-132, 140,186-187,197-201, 206,219-220,229-237

Here, for example, C^- represents the sum of the *percentage* compositions of His, Asp, and Asn in the protein of interest. The percentages of the four types of secondary structure are then calculated from the relations:

$$\% \text{ helix} = 16.37 + 1.582\ H^+ - 2.364\ H^- + \text{Heme} \quad [1]$$

$$\% \ \beta\text{-sheet} = 25.85 + 2.361\ B^+ - 1.917\ B^- \quad [2]$$

$$\% \text{ turn} = 43.17 + 0.896\ T^+ - 1.304\ T^- \quad [3]$$

$$\% \text{ coil} = 24.19 + 1.125\ C^+ - 2.041\ C^- \quad [4]$$

These coefficients were determined from the full data set of 18 proteins. As explained above, Heme = 20.00 for oxygen-carrying heme proteins and zero otherwise.

In order to test the validity of the procedure, we redetermined the coefficients in Eqs. 1–4, omitting one protein at a time; these were used to predict the percentage of helix, β-sheet, turns, and coil regions in the 18 proteins. The results obtained on omitting the protein to be predicted are displayed in Fig. 1. The line drawn with a 45° slope would represent perfect agreement between the predicted and assigned percentages. In Table 2 are listed the 18 proteins, the percentages assigned to the four types of structural features, and (under the four columns bearing the heading *Error 1*) the differences between the predicted and assigned percentages. As shown at the bottom of the table, for the 20 chains or subchains these errors average ±7.1% for helical content, ±6.9% for β-sheet, ±4.2% for turns, and ±5.7% for coil. These average errors are about the magnitude of the estimated uncertainty in the assignments of secondary structures in the initial data

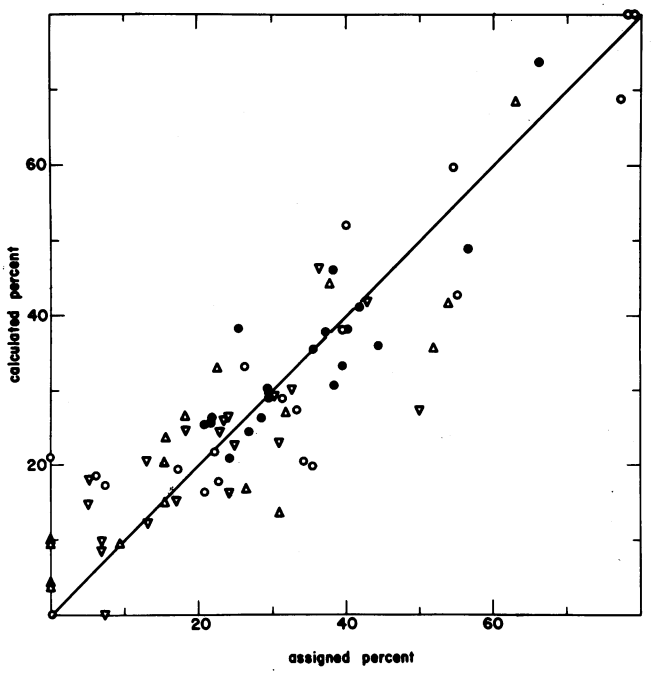

FIG. 1. Comparison of the predicted and assigned percentages of secondary structure for 18 proteins. O, Helix; △, β-sheet; ●, turn; ▽, coil.

TABLE 3. *Predicted amounts of secondary structure in other proteins*

	% Helix	% β-Sheet	% Turn	% Coil
L7/L12 ribosomal proteins	76	0	11	41
Aspartate transcarbamoylase (R-chain *Escherichia coli*)	44	17	30	10
α-Lactalbumin	32	35	49	4
Glyceraldehyde 3-phosphate dehydrogenase (pig)	29	33	37	17
Human carbonic anhydrase *c*	28	18	37	7
f_2 phage coat protein	24	22	28	27
Eylar basic protein	14	29	20	15
Immunoglobulin Γ-1 chain V-I, human EU	4	39	31	27
Immunoglobulin, κ chain V-I, human EU	4	43	26	30
Erythrocruorin IIβ (1–130)*	64	7	28	21
Erythrocruorin IVα (1–130)*	63	8	29	20
Lactate dehydrogenase*	33	26	46	0
Bovine superoxide dismutase*	25	22	46	9
Flavodoxin*	21	12	52	13
Bence–Jones MCG*	14	25	32	35

* Values calculated from the second set of relations.

set. Use of the coefficients given in Eqs. 1–4, which were determined from all 18 proteins, only reduced the average errors by about 15% of the values quoted above. This result would suggest that the procedure is not especially sensitive to exclusion of the protein being predicted from the data set; however, this conclusion must be regarded as tentative due to the presence of homologies within the data set.

The amount of secondary structure has been estimated experimentally by circular dichroism, optical rotatory dispersion, and infrared spectra. The uncertainties in these methods appear to be 5–10% in the best cases; however, none of the experimental methods allows a measure of all four categories predicted by the procedure described above.

The combinations of amino acids whose compositions yielded the best predictions of the percent helix, β-sheet, and turns require a distinction between the pairs Asp and Asn, and Glu and Gln. Often this information is not available, the residues simply being designated as Asx or Glx. We therefore

TABLE 2. *Assigned percentages of secondary structure and errors in prediction*

Protein	Helix			β-Sheet			Turns			Coil	
	Assigned	Error 1	Error 2	Assigned	Error 1	Error 2	Assigned	Error 1	Error 2	Assigned	Error 1
Myoglobin	79.1	1.0	-2.1	0.0	0.0	0.0	21.6	4.3	8.0	5.2	12.8
Oxyhemoglobin B	78.1	2.1	-8.8	0.0	3.9	9.0	29.4	0.5	1.4	6.9	1.7
Oxyhemoglobin A	77.3	-8.4	-7.1	0.0	10.1	16.7	28.4	-2.0	-11.4	5.0	9.9
Cytochrome b_5	55.2	-12.4	-20.3	26.4	-9.7	-15.9	25.3	13.1	18.6	6.9	3.0
Carp myogen	54.6	5.1	12.9	0.0	4.1	0.0	24.1	-3.0	3.3	24.1	-7.8
Insulin B	40.0	12.2	10.2	0.0	0.0	0.0	26.7	-2.0	-8.6	50.0	-22.6
Insulin A	33.3	-5.9	0.9	0.0	9.6	1.8	38.1	8.2	-4.4	42.9	-0.9
Lysozyme	39.5	-1.3	-14.2	15.5	5.0	23.1	39.5	-6.0	-5.0	17.0	-1.7
Carboxypeptidase	35.5	-15.6	-13.9	18.2	8.4	6.2	29.3	1.0	4.6	22.8	1.8
Thermolysin	34.2	-13.6	-9.6	22.5	10.7	6.2	41.8	-0.4	0.0	13.0	7.6
Subtilisin	31.3	-2.2	0.4	9.4	0.0	8.4	29.4	-0.2	-0.1	30.9	-7.9
Staph. nuclease	26.2	7.1	4.6	15.4	-0.4	-0.7	44.3	-8.0	-0.4	24.8	-2.1
Ribonuclease	22.6	-4.8	-0.7	37.9	6.6	2.8	21.8	4.6	7.2	23.4	2.8
Cytochrome *c*	22.1	-0.3	0.7	0.0	4.1	11.5	56.7	-7.5	-7.2	24.0	2.7
Papain	20.8	-4.3	-1.8	15.6	8.2	4.6	40.1	-1.7	0.8	30.2	-0.9
Trypsin inhibitor	17.2	2.2	0.6	31.0	-17.4	-1.9	20.7	4.9	7.2	36.2	10.2
α-Chymotrypsin	7.4	9.9	14.2	31.8	-4.7	-6.1	35.5	0.1	-3.3	32.6	-2.5
Elastase	6.3	12.4	13.1	52.1	-16.4	-24.9	37.1	1.0	1.2	18.3	6.5
Rubredoxin	0.0	0.0	0.0	63.0	5.5	-6.1	66.7	7.1	-10.1	7.4	-7.4
Concanavalin A	0.0	21.0	25.2	54.0	-12.3	-18.9	38.4	7.7	-6.5	13.1	-0.8
Average error		±7.1	8.1		±6.9	±8.2		±4.2	±5.5		5.7

repeated the procedure to obtain a second set of predictive functions that could be applied under these circumstances. In two cases the number of amino acids whose compositions are summed was increased beyond 5, and two regression relations were averaged to improve the prediction of percent helix. The second set of aminoacid combinations is:

helix: HA^+ = Ala + Leu + His + Ile ($r = +0.76$)
HA^- = Pro + Thr + Trp + Met + Tyr ($r = -0.76$)
HB^+ = Ala + Leu + His + Tyr ($r = +0.74$)
HB^- = Pro + Thr + Ile + Val + Arg ($r = -0.77$)

β-sheet: B^+ = Asp + Thr + Arg + Pro + Val + Asn ($r = +0.79$)
B^- = Leu + Ala + Glu + Gln + Gly ($r = -0.78$)

turn: T^+ = Gly + Thr + Asp + Gln + Glu + Met + Asn ($r = +0.79$)
T^- = Ser + Ala + Phe + His + Cys ($r = -0.81$)

and the corresponding relationships are:

$$\% \text{ helix} = 43.02 + 0.707\ HA^+ + 0.676\ HB^+ - 1.223\ HA^- - 0.865\ HB^- + \text{Heme} \quad [5]$$

$$\% \ \beta\text{-sheet} = 19.13 + 1.633\ B^+ - 1.477\ B^- \quad [6]$$

$$\% \text{ turn} = 25.91 + 0.904\ T^+ - 0.909\ T^- \quad [7]$$

The second method was tested in the same manner as before (i.e., the coefficients were redetermined for each protein predicted, omitting that protein from the data set). The errors are listed in Table 2 under the column headings *Error 2*. The average errors are 8.1 for percent helix, 8.2% for β-sheet, and 5.5% for turns. The second method is inferior in all three cases; hence the first set of relations is preferred if the available information is sufficient to distinguish the acid and amide sidechains. Where this is not the case, the second set of relations should furnish an estimate of sufficient reliability for many purposes. Table 3 illustrates an application of these relations to 15 additional proteins for which the sequence or tertiary structure is unknown. We suspect that the average errors for these proteins will be somewhat larger than the values given at the bottom of Table 2. Nevertheless, these predictions do furnish some interesting insights. The high predicted helical content of the L7/L12 ribosomal proteins is particularly noteworthy. 43% of the residues fall in the positive helical category, while only 5% correlate negatively.

We believe Eqs. **1–7** represent the first relations proposed for predicting the percentages of secondary structure from amino-acid composition. Troitskii and Zav'yalov (12) recently reported sets of amino acids whose composition correlated with the percentages of helix and β-sheet as estimated from optical rotatory dispersion data; however, they used this information to designate single residue potentials for predicting regions of secondary structure in proteins of known sequence. Further, their combinations of amino acids differ substantially from those given above.

The method proposed here is certainly capable of further refinement to yield more reliable predictions. For example, averaging of more than two regression relations (as was done in obtaining Eq. **5**) may give more reliable predictions. Secondly, as mentioned above, several aminoacid combinations gave high correlations (both positive and negative) for the four types of secondary structure. It is quite likely that stronger correlations will appear as additional protein structures are determined.

We thank Mr. R. Britton Mayo for assistance in programming the statistical analysis, and express our appreciation to Duke University for providing funds for the computer calculations. This work was supported in part by Grant GB40056 from the National Science Foundation.

1. Rubin, B. & Richardson, J. S. (1972) *Biopolymers* **11**, 2381–2385.
2. Venkatachalam, C. M. (1968) *Biopolymers* **6**, 1425–1436.
3. Matthews, B. W. (1972) *Macromolecules* **5**, 818–819.
4. Ptitsyn, O. B. & Finkelstein, A. V. (1970) *Biophysics* **15**, 785–796.
5. Leberman, R. (1971) *J. Mol. Biol.* **55**, 23–30.
6. Lewis, P. N. & Scheraga, H. A. (1971) *Arch. Biochem. Biophys.* **144**, 576–583.
7. Robson, R. & Pain, R. H. (1971) *J. Mol. Biol.* **58**, 237–256.
8. Harrison, S. C. & Blout, E. R. (1965) *J. Biol. Chem.* **240**, 299–303.
9. Breslow, E., Beychock, S., Hardman, K. D. & Gurd, F. R. N. (1965) *J. Biol. Chem.* **240**, 304–309.
10. Epand, R. M. & Scheraga, H. A. (1968) *Biochemistry* **7**, 2864–2872.
11. Hermans, J., Jr. & Puett, D. (1971) *Biopolymers* **10**, 895–914.
12. Troitskii, G. V. & Zav'yalov, V. P. (1972) *Mol. Biol.* **6**, 509–519.