Prediction of the Secondary Structure of Proteins Using the Helix–Coil Transition Theory[1]

Mark Froimowitz and Gerald D. Fasman*

Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02154.

Received April 17, 1973

ABSTRACT: Prediction of the secondary structure of proteins has been attempted using helix–coil transition theory. Values for use in the Zimm–Bragg helix–coil transition theory have been obtained from a statistical analysis of 15 proteins whose amino acid sequences are known and whose crystal structures have been determined. Utilizing new empirical values for these parameters makes it possible to obtain improved results over previous attempts. In addition, this study points out the main difficulty of this approach in that it omits the β structure as a possible state of the amino acid residues since many predicted helical regions are experimentally observed to be part of β regions.

The application of a matrix representation of the partition function of polypeptides to obtain the probability of an amino acid being in a particular state has been carried out by Zimm and Bragg[2] and Lifson and Roig.[3] More recently Scheraga and coworkers[4,5] have utilized the Zimm–Bragg formulation[2] of the same theory for predicting the helical regions of proteins from their amino acid sequences. However, this attempt was hampered, at least in part, by the lack of knowledge of the Zimm–Bragg parameters, s and σ, for all of the amino acids. To overcome this difficulty, the amino acids were grouped into three categories of helix breakers, helix indifferent, and helix formers. Each of the categories was then *tentatively* assigned a value of s pending the outcome of their experimental determination.[6] In addition, a single value of $\sigma = 5 \times 10^{-4}$ was assigned to all amino acids.

Since it now is possible to obtain a set of empirical values, using statistical data from the experimental X-ray crystal structures of proteins,[7] which can be used in lieu of s and σ, this present study was initiated to see if it were possible to make improved predictions of helicity for proteins using the above theories. To this end, calculations have been performed on 15 proteins whose amino acid sequences are known and whose crystal structures have been determined by X-ray crystallography.

An additional problem encountered by Scheraga and coworkers was the lack of an appropriate criterion for deciding whether a residue was in a helical region, since the helical probabilities that they calculated for the residues never exceeded several per cent. Thus, the calculated probability of being in the coil was always over 90% for all of the residues. Therefore, they adopted the *average* calculated helical probability of the protein under consideration as a cutoff point. Any residue in a protein whose calculated helical probability was greater than the average for the protein was predicted to be helical. However, using an average helical probability would be expected to give poor results for proteins with small or large amounts of helix. For polypeptides such as the A and B chains of chymotrypsin, which do not contain any helix at all, it is obvious that poor predictions will be made based on this criterion.

To partially overcome this problem, a different cut-off point for predicting helical residues has been adopted. Instead of the average calculated helical probability, the *experimentally observed* helix content has been substituted as the cut-off point. When a protein is found to have a certain helical content by experimental means such as optical rotatory dispersion or circular dichroism, it is predicted that the same number of residues with the *largest* calculated helical probabilities would be contained in the helical regions. While it is not always possible to accurately measure the helicity of a protein using the above experimental methods, for most proteins it is possible to obtain reasonable estimates.[8,9]

From the point of view of this study, the per cent of experimentally observed helix is useful for yet another reason. Since examination will be made of proteins whose crystal structures are known, it will be easy to determine the exact number of helical residues present in the protein. One would now want those experimentally observed helical residues to have the largest calculated helical probabilities, and if they do not, one would try to determine the reason for this discrepancy. This procedure should allow one to examine the limitations and shortcomings of the model used in this study for the purpose of predicting the secondary structure of proteins.

Theory

In this paper, utilization of the Lifson–Roig formulation[3] rather than that of Zimm–Bragg[2] has been made for reasons that will become evident in what follows. Using this model, the partition function of a polypeptide is

$$Z = (0,0,1)\left[\overset{N}{\underset{i\,=\,1}{\pi}} \quad \mathbf{W}^{j}(i)\right]\begin{pmatrix}0\\1\\1\end{pmatrix} \qquad (1)$$

where $\mathbf{W}^{j}(i)$ is a 3×3 matrix of statistical weights for the ith residue in the polypeptide chain which is of amino acid type j, N is the number of residues in the chain, and where the indicted matrix multiplications are performed. In an effort to make the model more realistic for heteropolypeptides, minor modifications have been made in the assignment of statistical weights in matrix $\mathbf{W}^{j}(i)$. In their formulation, Lifson and Roig[3] assigned a single statistical weight, v, to all helical residues that occur at a helix-coil boundary. In this work, however, it was decided that the v for a residue should differ depending on whether it is at the amino or carboxyl end of a helical section. In actual fact, it appears that certain amino acids have a tendency to appear at one or another end of a helical segment.[7,10-12] Making this differentiation, the matrix becomes[13]

$$\mathbf{W}^{j}(i) = \begin{pmatrix} w^{j}(i) & v_{c}^{\,j}(i) & 0 \\ 0 & 0 & 1 \\ v_{n}^{\,j}(i) & v_{n}^{\,j}(i)\,v_{c}^{\,j}(i) & 1 \end{pmatrix} \qquad (2)$$

with the statistical weights assigned as follows: w for the addition of a helical residue to a long helical section, v_n and v_c for residues that are helical but occur respectively at the amino and carboxyl ends of a helical section, $v_n v_c$ for an isolated helical residue,[14] and 1 for a residue in the coil region. (Since most of the literature uses the Zimm–Bragg notation of s and σ, we will continue to use both notations. The relationships between the notations are $s = w$ and $\sigma = v_n v_c$.)

Given the above partition function, the probability that the ith amino acid will be helical is[3]

$$P_H(i) = (0,0,1) \times \left[\begin{matrix} i-1 & \\ \pi & \mathbf{W}^j(k) \\ k=1 & \end{matrix}\right] \begin{pmatrix} w^j(i) & v_c{}^j(i) & 0 \\ 0 & 0 & 0 \\ v_n{}^j(i) & v_n{}^j(i) v_c{}^j(i) & 0 \end{pmatrix}$$

$$\left[\begin{matrix} N & \\ \pi & \mathbf{W}^j(k) \\ k = i-1 & \end{matrix}\right] \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} \Big/ Z \qquad (3)$$

where all the terms have been defined above. Equation 3 is to be interpreted as calculating the contribution to the partition function of the helical states of residue i (while omitting its coil states) where one is averging over all possible states of the remaining amino acid residues.

The reason for choosing the Lifson–Roig formulation rather than that of Zimm–Bragg should now be apparent. In the latter treatment, a single statistical weight σ, which is equivalent to $v_n v_c$, is assigned to the beginning of a helical section. While this may be adequate for a homopolymer where all residues are identical, it would not be realistic for a protein having many different amino acids since each helical segment of the protein would probably have different amino acids at either end of the helical segment.

The statistical weights, v_c and v_n, that have been assigned to the ends of a helical segment have to be interpreted in a slightly different manner than their usual meaning as a nucleation parameter for the formation of helices. When an amino acid residue is incorporated into a helical segment, there are two conflicting terms in its free energy of interaction. There is a decrease in the entropy of the residue due to its restriction to a smaller volume of conformation space and there is a decrease in the energy due to the stabilization by hydrogen bonds. The first of these terms leads to a lowering of the statistical weight of the helical state whereas the second leads to an increase of the statistical weight. If one now examines a sufficiently long helical segment (seven or more helical residues), one notices that the central residues will each have two hydrogen bonds whereas the three residues at either end will only have a single hydrogen bond. Therefore, one would expect these end helical segments to have lower statistical weights than the central residues since they are being stabilized by one less hydrogen bond while, at the same time, having restricted conformation spaces. Similar reasoning was used in the assignment of statistical weights to amino acid residue states in a more detailed model of the helix–coil transition that has been proposed for specific sequence polypeptides.[15] Though it would be more correct to spread out this destabilization over the statistical weights of the three end residues of helical segments, it has only been assigned to the single end residue since it does not seem to change the overall probabilities very much.[15] The above interpretation of v_c and v_n is, of course, compatible with the nucleation behavior found in the helix–coil transition of polypeptides.

An additional point that must be discussed is the applicability of the above model to the three-dimensional structure of proteins. It is clear that helix–coil transition theory only concerns itself with the short-range interactions that occur along the one-dimensional polypeptide axis and there is no consideration at all of long-range interactions. While there are certainly long-range interactions occurring in a native globular protein, there is a great deal of evidence that the secondary structure of proteins is predominantly determined by the short-range interactions of the amino acid sequence. Otherwise, one certainly could not attempt to predict the secondary structure of proteins from the primary sequence in the manner that many researchers, with varying degrees of success, are doing.[16] Indeed, the results of this paper would lead one to make a similar conclusion.

Table I
Parameters Used in Calculations

Amino acid	Tentative s used previously[a]	s' used in this paper[b]	$v_n{}'$ [b,c]	$v_c{}'$ [b,c]	$\sigma' \times 10^3$ [d]
Ala	1.05	1.63	0.070	0.059	4.1
Arg	1.00	0.51	0.019	0.077	1.5
Asn	0.385	0.43	0.045	0.045	2.0
Asp	1.00	0.68	0.104	0.027	2.8
Cys	1.00	0.56	0.037	0.074	2.7
Gln	1.05	1.29	0.058	0.079	4.6
Glu	1.05	1.35	0.098	0.062	6.1
Gly	0.385	0.29	0.030	0.020	0.60
His	1.05	1.03	0.027	0.108	2.9
Ile	1.05	1.03	0.033	0.038	1.3
Leu	1.05	1.54	0.028	0.051	1.4
Lys	1.00	0.75	0.029	0.080	2.3
Met	1.05	1.50	0.036	0.072	2.6
Phe	1.00	1.03	0.049	0.055	2.7
Pro	0.385	0.00	0.106	0.000	0.0
Ser	0.385	0.50	0.040	0.042	1.7
Thr	1.00	0.63	0.061	0.023	1.4
Trp	1.05	1.06	0.068	0.023	1.6
Tyr	1.00	0.41	0.035	0.025	0.88
Val	1.05	1.30	0.031	0.058	1.8

[a] Reference 5. [b] Obtained from ref 7. [c] $v_n{}' = 0.50 f_{hn}$, $v_c{}' = 0.05 f_{hc}$, $\sigma' = 0.25 f_{hn} f_{hc}$ (see text). [d] $\sigma' = v_n{}' v_c{}'$.

Parameters

The replacement values for s and σ that have been used herein (s' and σ') are listed in Table I. With a single exception, they were obtained from a statistical analysis of 15 proteins whose amino acid sequence and structure in the crystal have been determined.[7,17]

The values used for s' were taken from the relative occurrence of the various amino acids in their helical and coil states in the above set of proteins.[7] (These numbers are reproduced in Table V.) In addition, a rationale for following this procedure is discussed in the Appendix. The other parameters listed in Chou and Fasman (Table II), P_α and $P_{\alpha i}$, were also utilized for calculations, but the results with the former parameters gave somewhat better results. The reason for this seemed to be that the values of n_α/n_c for the helix breakers tended to be much smaller than those of the other parameters (P_α and $P_{\alpha i}$) and this resulted in the helical segments being terminated much more in line with the experimental X-ray data. The only change from Chou and Fasman[7] introduced here is to set the value of s' for proline to zero. This is a reasonable change as it is well established that proline cannot be incorporated into a helix without disrupting it.

The values for v_n and v_c were obtained from the f_{hn} and the f_{hc} of Table IV (ref 7), where f_{hn} and f_{hc} are defined as the fraction of residues at the amino and carboxyl ends, respectively, of helical segments in proteins. To a first approximation, v_n and v_c would be expected to be proportional to the above fractions since the probability of an amino acid occurring in a particular state would be expected to be proportional to that state's statistical weight. The

Table II
Number of Residues Wrongly Predicted Using Different Sets of Parameters

	1	2	3	4	5
Protein	% helix from X-ray	Tentative s[a] $\sigma = 5 \times 10^{-4}$	Statistical s'[b] $\sigma = 5 \times 10^{-4}$	Statistical s'[b] variable σ'[b]	Statistical s'[b] variable σ'[b] av helical probability cut off[a,c]
Carboxypeptidase[d]	38	84	64	66	56
Chymotrypsin, C chain[e]	27	22	22	22	23
Cytochrome b_5[f]	50	42	34	26	26
Cytochrome c[g]	39	30	26	26	29
Elastase[h]	7	34	34	34	85
α-Hemoglobin[i]	77	44	28	28	48
β-Hemoglobin[i]	78	40	40	32	57
Insulin, A chain[j]	67	6	2	2	6
Insulin, B chain[j]	37	2	0	0	0
Lysozyme[k]	46	26	20	18	18
Myogen[l]	48	54	54	46	52
Myoglobin[m,r]	79	42	36	34	68
Papain[n]	26	46	42	40	41
Ribonuclease A[o]	26	18	16	18	34
Staphylococcal nuclease[p]	24	56	24	18	26
Subtilisin BPN[q]	31	80	80	76	75
Total		626	522	486	644

[a] Reference 5. [b] Reference 7. [c] Reference 4. [d] Reference 19. [e] J. J. Birktoft, *et al., Phil. Trans. Roy. Soc. London, Ser. B*, **257,** 67 (1970). [f] F. S. Mathews, *et al., Cold Spring Harbor Symp. Quant. Biol.*, **36,** 387 (1971). [g] R. E. Dickerson, *et al., J. Biol. Chem.*, **246,** 1511 (1971). [h] D. M. Shotton and H. C. Watson, *Nature (London)*, **225,** 811 (1970). [i] M. F. Perutz, *et al., Nature (London)*, **219,** 131 (1968). [j] T. Blundell, *et al., Cold Spring Harbor Symp. Quant. Biol.*, **36,** 233 (1971). [k] C. C. F. Blake, *et al. Proc. Roy. Soc., Ser. B*, **167,** 365 (1967); C. C. F. Blake, *et al., Nature (London)*, **206,** 757 (1965). [l] C. E. Nockolds, *et al., Prcc. Nat. Acad. Sci., U. S.*, **69,** 581 (1972). [m] H. C. Watson, *Progr. Stereochem.*, **4,** 299 (1969). [n] B. C. Wolthers, *et al.*, "Structure-Function Relationships of Protolytic Enzymes," P. Desnuelle, H. Neurath, and M. Ottesen, Ed., Academic Press New York, N. Y., 1969, p 272. [o] G. Kartha, *et al., Nature (London)*, **213,** 862 (1967). [p] A. Arnone, *et al., J. Biol. Chem.*, **246,** 2302 (1971). [q] R. A. Alden, *et al., Phil. Trans. Roy. Soc. London, Ser. B*, **257,** 119 (1970). [r] The criteria listed in the parameter section have not been used for determining whether a residue is helical or not for myoglobin, since by those criteria only 5 out of the 153 residues are not helical. Instead, the reported results of the author have been used, which include only α helices.

Table III
Results Showing the Number of Residues and Helical Regions Predicted Correctly

Protein[a]	No. of residues	Residues predicted correctly	% of residues predicted correctly	% of helical residues predicted correctly	Helical regions predicted correctly	Regions predicted to be helical which are not
Carboxypeptidase	307	241	79	75	6/8	3
Chymotrypsin, C chain	97	75	77	58	1/2	2
Cytochrome b_5	93	67	72	70	5/6	1
Cytochrome c	104	78	75	68	4/6	0
Elastase	240	206	86	0	0/2	2
α-Haemoglobin	141	113	80	87	6/7	0
β-Haemoglobin	146	114	78	85	7/8	0
Insulin, A chain	21	19	90	93	2/2	0
Insulin, B chain	30	30	100	100	1/1	0
Lysozyme	129	111	86	85	6/6	1
Myogen	108	62	57	56	4/6	0
Myoglobin	153	119	78	86	8/8	0
Papain	212	172	81	73	4/5	1
Ribonuclease A	124	106	85	63	2/3	0
Staphylococcal nuclease	149	131	88	74	3/3	0
Subtilisin BPN′	275	199	72	56	6/8	4
Total	2329	1843	79	74	65/81	14

[a] See Table II for sources of X-ray data.

values for the above parameters that are listed in Table I (herein) include a proportionality factor of 0.25 (*i.e.*, $v_n = 0.50 f_{hn}$, $v_c = 0.50 f_{hc}$, $\sigma = 0.25 f_{hn} f_{hc}$) since that proportionality factor gave the best agreement with the X-ray data although factors of 0.50 and 0.10 gave results that were almost equivalent.

The manner in which values utilized for s and σ have been obtained for this paper is quite different from their usual experimental determination with copolymers.[6,18] Since the values from the above parameters listed in Table I have been obtained from the crystal structures of native proteins, they are *not* equivalent to the s and σ values obtained from copolymer studies, since the former will contain some contribution from long-range interactions of the folded polypeptide chain and also from the internal hydrophobic environment that is characteristic of proteins. For that reason, the values that are listed in Table I have been denoted as s', v_n', and v_c' since they are to be distinguished

Table IV
Comparison of Predicted Helical Regions with Experimentally Observed α and β Regions for Proteins

Protein[a]	X-Ray resolution[b] Å	Predicted[c]	Experimentally obsd α Helix	Experimentally obsd β Region[d]	Comments[e]
Carboxypeptidase	2.0	14–29	14–28		
		31–37		32–36	
		63–69		60–67	
		72–85	72–88		
		97–110	94–103		
			112–122		6/11 have $s' < 1.0$
		172–185	173–187		
		189–195		190–196	
		215–233	215–231		
			254–262		6/9 have $s' < 1.0$
		289–305	285–306		
Chymotrypsin,[f] C chain	2.0	6–9		7–15	
			16–25		9/10 have $s' < 1.0$
		30–34		31–35	
		80–96	82–97		
Cytochrome b_5	2.0	8–15	8–15		
		21–27		21–25	
		31–39	33–38		
		43–50	42–49		
		54–61	55–62		
		66–71	64–74		
			80–86		6/7 have $s' < 1.0$
Cytochrome c	2.8	7–21	9–13, 14–18		5/6 have $s' < 1.0$
			49–54		
		58–69	62–70		
			71–75		4/5 have $s' < 1.0$
		89–100	91–101		
Elastase[f]	3.5	40–46		38–44	
		93–101		93–101	
			154–160		5/7 have $s' < 1.0$
			231–240		4/9 have $s' < 1.0$
α-Haemoglobin	2.8	3–17	3–18		
		23–35	20–35		
			36–42		6/7 have $s' < 1.0$
		45–73	52–71		
		79–92	80–88		
		95–112	94–112		
		119–137	118–138		
β-Haemoglobin	2.8	2–34	4–18, 19–34		
			35–41		4/7 have $s' < 1.0$
		52–55	50–56		
		59–69	57–76		
		73–97	85–93		
		101–118	99–117		
		124–145	123–143		
Insulin, A chain	2.8	2–7	2–8		
		11–18	12–18		
Insulin, B chain	2.8	9–19	9–19		
Lysozyme	2.0	4–15	4–15		
		27–37	24–36		
		40–42		41–54	
		81–85	80–85		
		87–97	88–100		
		105–113	108–115		
		119–125	119–125		
Myogen	2.0	8–31	7–15, 26–33		
		42–51	40–51		
		58–75	67–71		
			78–89		8/12 have $s' < 1.0$
			102–107		2/6 have $s' < 1.0$
Myoglobin	1.4	3–33	3–18, 20–35		
		38–86	36–42, 51–57, 58–77		
		89–93	86–95		
		103–117	100–118		
		125–145	124–149		
Papain	2.8	26–38	24–43		
		49–56	50–58		
		68–78	69–78		
			116–126		5/11 have $s' < 1.0$
		129–143	138–143		
		156–164		162–175	
Ribonuclease A	2.0	3–11	5–12		
			28–35		4/8 have $s' < 1.0$
		46–60	51–58		

Table IV (continued)

Protein[a]	X-Ray resolution[b] Å	Predicted[c]	Experimentally obsd α Helix	β Region[d]	Comments[e]
Staphylococcal nuclease	2.0	60–74	54–67		
		101–105	99–106		
		122–136	122–134		
Subtilisin BPN′	2.5		5–10		4/6 have $s' < 1.0$
		15–17	14–20		
		27–30		28–32	
		69–76	64–73		
		89–95		89–94	
		111–122	103–117		
		131–155	132–145		
		175–177			
		195–199			
		225–237	223–238		
			242–252		7/11 have $s' < 1.0$
		269–274	269–275		

[a] See Table II for sources of X-ray data. [b] One should take the resolution into account in evaluating the X-ray results for a protein, since at poorer resolutions there is much more uncertainty in the locations of the atoms. [c] Several predicted segments that are only one or two residues long have been omitted. [d] Only those β regions which coincide with predicted helical regions have been listed. [e] If the value of s' is greater than one for an amino acid, it would tend to be a helix former. The converse is also true. [f] The numbering system used by the original authors has not been followed. Instead, the residues have been numbered consecutively from the amino end of the protein.

from the usual definition of s and σ which would not contain the above protein effects.

In addition, an assumption that has been made is that each amino acid residue has a single value of s, v_n, and v_c *independent* of the residues that are contiguous to it. While this is certainly an approximation, experimentally measured values of s and σ suffer from the same deficiency since the experimentally measured values of these parameters will presumably depend on neighboring residues. Indeed, it appears that the values of the above parameters obtained from proteins might actually be more germane to protein structure since they will have been obtained from the *averaged* environment that is present in proteins.

The per cent helicity observed in the crystal was utilized as the criterion for predicting whether a residue was to be found in a helical region for the reasons outlined in the introductory section. If a certain number of residues are observed to be helical in the crystal, the same number of residues which have the largest calculated helical probabilities are predicted to be helical.

It is occasionally difficult to decide which residues should be included in a helical region from the reported X-ray crystallographic results of some proteins. In an effort to be consistent, which is not the case between different groups reporting X-ray structures, any residue was considered to be helical if it satisfied either of the following two criteria: (1) it contributed a hydrogen bond to a helix, or (2) while not contributing a hydrogen bond, it was in the center of a helical region with the appropriate (ψ, Ψ) angles. That these are different criteria is best illustrated by the review of the X-ray study of carboxypeptidase[19] where the authors report that residues 112–122 are helical though they contain only one or two hydrogen bonds. In addition, all variations of the α helix (α_{11}, 3.0_{10}, etc.) have been considered to be helical. Following this procedure has led to defining some residues to be helical though the original authors may not have done so. Finally, an attempt has been made to utilize the latest revisions of the X-ray results.

Results

Before comparing the predictions of the secondary structure of several proteins with their crystal structures, it would be interesting to evaluate the improvement brought about by using the values of s' and σ' obtained from the statistical analyses of proteins.[7] The results of calculations using various sets of parameters are compiled in Table II, where the number of residues predicted *incorrectly* for each set are listed. One sees that there is a significant drop in the number of incorrectly predicted residues with the new values of s' (using $\sigma = 5 \times 10^{-4}$) over the tentatively assigned values used previously (compare columns 2 and 3).[5] There is an additional improvement when one also uses the new values of σ' (compare column 3 and column 4).

Also listed in Table II are the results obtained by using the average helical probability cut-off point used by Scheraga and coworkers (column 5).[4,5] As was discussed in the introductory section, this criterion gives poor results for those proteins that have large or small amounts of helix such as myoglobin, the haemoglobins, and elastase. For the other proteins under consideration, the results are very similar to those obtained by using the experimentally observed helicity.

More complete results, using the parameter set of Table I, are presented in Tables III and IV. In Table III are listed the percentage of residues and helical residues predicted correctly. Overall, 79% of all residues have been predicted correctly and 74% of all helical residues predicted correctly. (Both numbers should be used together as can be seen by the result for elastase.) In addition, the results have been interpreted in terms of helical regions predicted correctly. A helical region is considered to be correctly predicted if there are three or more predicted residues coincident with an experimentally observed helical region. Using this criterion, 80% of the 81 helical sections present are correctly predicted in these 15 proteins. The actual results, however, are much better than this rather broad criterion might indicate. One finds (Table IV) that most of the correctly predicted helical regions almost coincide with the regions found in the crystal by X-ray crystallography.

In addition to the correctly predicted helical regions, 14 extra regions are predicted to be helical that are not experimentally observed to be so (Table IV). However, it is seen that 12 of these 14 predicted regions are experimentally observed to be part of β regions. The fact that this occurred points out one of the limitations of the model that has been chosen. In addition to the helical and coil states that the

Table V
Relative Occurrence of Helical and Coil States for Amino Acids in a Set of Proteins[a]

Amino acid	n_α/n_c[b]	Amino acid	n_α/n_c[b]
Ala	1.68	Leu	1.54
Arg	0.50	Lys	0.78
Asn	0.42	Met	1.50
Asp	0.68	Phe	1.06
Cys	0.56	Pro	0.31
Gln	1.14	Ser	0.48
Glu	1.35	Thr	0.61
Gly	0.29	Trp	1.06
His	1.03	Tyr	0.39
Ile	0.97	Val	1.32

[a] Data are taken from ref 7. [b] n_α and n_c are respectively the number of residues in the helical and coil regions.

model considers, one must also include the β state since there seems to be some overlap between helix-forming and β-forming residues. Thus, the partition function in eq 1 is actually incomplete since the possibility that residues may become part of a β region has been omitted. A more complete treatment of the conformational states of proteins would necessitate the competition among all *three* possible states. Unfortunately, it is not at all clear at present how to incorporate the β state into the model. In fact, one would anticipate many difficulties in doing so since one is required to deal with the complex long-range interactions of *different* parts of the polypeptide chain or even of different polypeptide chains.[20]

Of the 16 experimentally observed helical regions that were not predicted, it can be noted that many of these contain a majority of helix-breaking residues and it is not apparent why they are helical. Of course, long-range interactions have not been taken into account, and it may turn out that such interactions stabilize these regions in a helical conformation. Teleologically, however, one wonders why these "nonhelical" residues were not replaced through the process of evolution. In addition, some of the missed regions seem to have been caused by overprediction of helix in other parts of the protein. Thus, if the β state had been included in the model, one would have predicted more helical regions correctly.

Conclusions

The main conclusion of this study is that one is able to make improved predictions of helicity over previous attempts using helix–coil transition theory. The previous work by Scheraga and coworkers[4,5] utilized tentatively assigned values of s and σ while the empirical set of values used for these parameters herein (s' and σ') was obtained from a statistical analysis of proteins with known amino acid sequences and crystal structures.[7] Overall, it has been possible to correctly predict 79% of all residues and 74% of all helical residues. In addition, the prediction of 80% of all helical regions was feasible.

This study has also illustrated the main shortcoming of the helix–coil transition model when it is used for making predictions of the secondary structure of proteins. That is, one cannot ignore the β conformation as one of the possible states for amino acid residues in a protein. This omission results in many experimentally observed β regions of proteins being predicted as α helical. This criticism would, of course, also apply to a more detailed model for the helix–coil transition which was developed for specific sequence polypeptides.[15]

Appendix I

The values of s' listed in Table I have been obtained by counting the relative occurrence of the various amino acids in their helical and coil states in a set of crystalline proteins whose conformations have been determined by X-ray crystallography. An analogy can be made to equilibrium constants. The equilibrium constant, $K_{c\rightarrow\alpha}$, for an amino acid in equilibrium with its coil and helical states is $K_{c\rightarrow\alpha} = n_\alpha/n_c$, where n_α and n_c are respectively the relative number of times that an amino acid is measured to be helical and coil states (Table V). In order for the above equation to be applicable to the amino acids in a crystallized protein, however, one must assume that the conformation of the protein in the crystal state is identical with its conformation in solution where an equilibrium state exists. This assumption is generally accepted and is indeed the justification for doing X-ray crystallography of proteins. Additionally, there is ample experimental evidence supporting the contention that the conformation of a protein is similar in the two states.[21,22]

A second and more questionable assumption that must be made is that the various amino acid residues are acting independently of each other. With this assumption, the Zimm–Bragg parameters will indeed be equal to $K_{c\rightarrow\alpha}$ and the parameter σ will be equal to 1. This second assumption is questionable since there is cooperativity of interaction along the polypeptide chain and there may also be long-range interactions occurring in proteins.

References and Notes

(1) This is Publication No. 959 from the Graduate Department of Biochemistry, Brandeis University. This research was generously supported by Grants from the U. S. Public Health Service (GM 17533), National Science Foundation (GB 29204), American Heart Association (71-111), and The American Cancer Society (P-577).

(2) B. H. Zimm and J. K. Bragg, *J. Chem. Phys.*, **31**, 526 (1959).

(3) S. Lifson and A. Roig, *J. Chem. Phys.*, **34**, 1963 (1961).

(4) P. N. Lewis, N. Gō, M. Gō, D. Kotelchuck, and H. A. Scheraga, *Proc. Nat. Acad. Sci. U. S.*, **65**, 810 (1970).

(5) P. N. Lewis and H. A. Scheraga, *Arch. Biochem. Biophys.*, **144**, 576 (1971).

(6) H. A. Scheraga, *Chem. Rev.*, **71**, 195 (1971).

(7) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 211, 222 (1974).

(8) A. J. Adler, N. Greenfield, and G. D. Fasman, *Methods Enzymology*, **27**, 675 (1973).

(9) Y. H. Chen, J. T. Yang, and H. M. Martinez, *Biochemistry*, **11**, 4120 (1972).

(10) A. V. Finkelstein and O. B. Ptitsyn, *J. Mol. Biol.*, **62**, 613 (1971).

(11) B. Robson and R. H. Pain, *Nature (London), New Biol.*, **238**, 107 (1972).

(12) J. L. Crawford, W. N. Lipscomb, and C. G. Schellman, *Proc. Nat. Acad. Sci. U. S.*, **70**, 538 (1973).

(13) This implies that one is enumerating the residues from the amino terminal end toward the carboxyl end.

(14) An isolated helical residue has a very specific meaning in the context of helix–coil theory. This type of residue is defined as having its conformation space *restricted* to the α_R region, yet at the same time having *no* hydrogen bonds. Thus, there is only an unfavorable entropy term for a residue in this state and one would, therefore, expect its statistical weight to be relatively small. In the above assignments, however, the statistical weight of the isolated helical residue has probably been underestimated since it has been assigned a statistical weight of $v_n v_c$. As has been pointed out by one of the reviewers, this is not completely realistic in a physical sense and also differs from the assignment of v to this state by Lifson and Roig. However, this would not affect the computational results, since the contribution of isolated helical residues, with their relatively low statistical weights, would not be significant to the overall helical probabilities.

(15) N. Gō, P. N. Lewis, M. Gō, and H. A. Scheraga. *Macromolecules*, **4**, 692 (1971).

(16) See ref 7 for more complete references to this type of work.

(17) There are small differences between the values listed in Table I of this paper and the values of n_α/n_c calculated from Table II of Chou and Fasman.[7] This is due to revisions in the assignment of various amino acid residues of the 15 proteins under consideration subsequent to the completion of the work of this paper. These minor changes, however, would not be expected to change the conclusions of this work.

(18) P. H. von Dreele, D. Poland, and H. A. Scheraga, *Macromolecules*, **4**, 396 (1971).

(19) F. A. Quiocho and W. N. Lipscomb. *Advan. Protein Chem.*, **25**, 1, (1971).

[20] After the completion of this work, the latest revision of the X-ray data on subtilisin BPN showed that the two regions 175–177 and 195–199 are also part of β regions (J. Kraut, private communication). Thus, of the 14 extra regions that are predicted to be helical, *all* are experimentally found to be part of β regions. This reinforces the conclusions of this paper.

[21] N. Yu and B. H. Jo, *Arch. Biochem. Biophys.*, **156**, 469 (1973).
[22] J. A. Rupley, *Biol. Macromol.*, **2**, 291 (1969).

Polybenzazoles Containing 2-Benzimidazolyl Side Groups

V. V. Korshak, E. S. Krongauz, A. P. Travnikova, and A. L. Rusanov*

Institute of Elementoorganic Compounds, Academy of Sciences of the USSR, Moscow, Vavilova 28, USSR. *Received August 14, 1974*

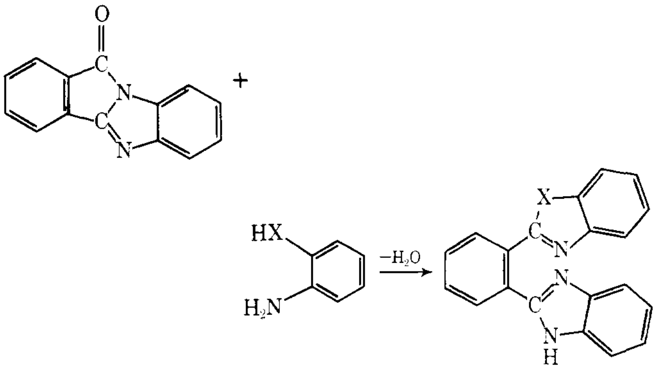
ABSTRACT: 1,2-Benzoylenebis(benzimidazoles) were allowed to react with tetrafunctional nucleophilic compounds [bis(*o*-phenylenediamines), bis(*o*-aminophenols), and bis(*o*-aminothiophenols)] in poly(phosphoric acid) (PPA) to prepare polybenzimidazoles (PBI), polybenzoxazoles (PBO), and polybenzthiazoles (PBT) containing 2-benzimidazolyl side groups. The introduction of 2-benzimidazolyl side groups leads to the formation of thermally stable polybenzazoles which are soluble in different organic solvents.

Introduction of thermally stable side groups in macromolecules is the most efficient route to improve solubility and, hence, tractability of rigid-chain polymers.[1] Thus, the introduction of phenyl substituents in polyphenylenes,[2] polyquinoxalines,[3,4] and polybenzimidazoles[5] considerably increases the solubility of these polymers; it should be noted that the introduction of nonpolar phenyl substituents decreases in most cases the softening temperatures of the polymers. Therefore, we have made an attempt to increase the solubility of the rigid-chain polymers by introduction of the polar 2-benzimidazolyl side groups in these macromolecules. Aromatic polybenzazoles (PBI, PBO, and PBT) containing 2-benzimidazolyl substituents have been studied for this purpose. These polymers were synthesized by the reaction of 1,2-benzoylenebis(benzimidazoles) with tetrafunctional nucleophilic compounds—bis(*o*-phenylenediamines), bis(*o*-aminophenols), and bis(*o*-aminothiophenols). The synthesis of polybenzazoles with 2-benzimidazolyl substituents was based on the ability of the CO—N< bond of the benzoylenebenzimidazole ring to cleave easily with *o*-phenylenediamine.[6,7]

Model Compounds

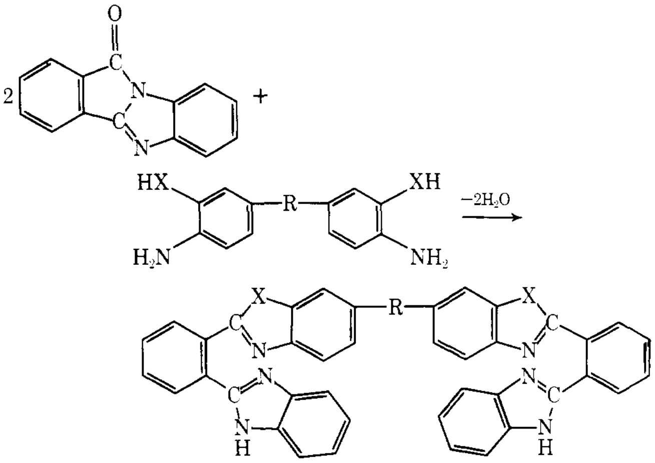
Prior to polymer synthesis model reactions and compounds were studied to determine the most favorable conditions for polymerization reactions and the structure of the resulting polymers. The model reactions were especially necessary, as the cleavage of 1,2-benzoylenebis(benzimidazole) was previously carried out only by treatment with *o*-phenylenediamine in the melt[6] while terephthaloylenebenzimidazole was cleaved with *o*-phenylenediamine in nitrobenzene.[7] Performance of this process in PPA showed that under these conditions a purer end product, *o*-phenylenebis(benzimidazole), is formed in almost quantitative yield. Since the high yields in these reactions make possible the preparation of high molecular weight polymers and since the polycyclocondensation in PPA is one of the most suitable methods for polyheteroarylene synthesis,[8,9] we have chosen PPA as the preferred reaction medium.

We extended the reaction of 1,2-benzoylenebis(benzimidazole) with *o*-phenylenediamines to other o-substituted anilines[10]—*o*-aminophenol and *o*-aminothiophenol—thus showing the general character of the reaction which may be represented by



X = NH, O, S

More complex model compounds were prepared by the reaction of 1,2-benzoylenebis(benzimidazole) with bis(*o*-phenylenediamines), bis(*o*-aminophenols), and bis(*o*-aminothiophenols)



X = NH, O, S; R = —O—, —CH₂—, —C(=O)—

On the other hand, bis-model compounds were synthesized by the reaction of 1,2-benzoylenebis(benzimidazoles) with *o*-phenylenediamine, *o*-aminophenol, and *o*-aminothiophenol.