A Model for the Helix–Coil Transition in Specific-Sequence Copolymers of Amino Acids[1]

Nobuhiro Gō,[2a] Peter N. Lewis,[2b] Mitiko Gō, and Harold A. Scheraga*

Department of Chemistry, Cornell University, Ithaca, New York 14850.

Received July 19, 1971

ABSTRACT: A model for the helix–coil transition in specific-sequence copolymers of amino acids is presented. While the statistical weight matrix, required for the correlation of the states of four consecutive residues in this model, is 44 × 44, it is shown that this can be contracted to an 11 × 11 matrix, whose secular equation is the same as that for the Zimm–Bragg 8 × 8 matrix, which is contractable to a 4 × 4 matrix. The statistical weight of a residue in one of eight distinct states reflects the major intramolecular interactions responsible for helix stabilization. It is shown that, for copolymers, the form of the individual statistical weights is unique, rather than arbitrary as it is for homopolymers. While the 11 × 11 matrix is required for a copolymer, it reduces to a 4 × 4 matrix for a homopolymer. Numerical examples on a specific-sequence binary random copolymer indicate that the average helix contents, as determined by the 11 × 11 and by the Zimm–Bragg 2 × 2 matrix methods, differ in most cases by only a few per cent. Thus, the 2 × 2 matrix formulation can be used as a good approximation for the analysis of experimental data on polypeptide copolymers. However, the 11 × 11 matrix formulation has much more information than the 2 × 2 matrix formulation about the conformational state of each residue in a specific-sequence random copolymer. The new formulation has been applied to sperm whale myoglobin and hen egg white lysozyme to demonstrate how the more precise specification of the state of an amino acid residue in a denatured protein allows interesting correlations to be made with the corresponding three-dimensional structure in the native conformation.

Statistical mechanical theories of the helix–coil transition in *homopolymers* of amino acids[3–6] were originally formulated, explicitly or implicitly, in terms of specific models for the intramolecular interactions (*e.g.*, the assumption of the domination of hydrogen-bonding interactions for the stabilization of the right-handed α helix). The transition curves were computed in terms of two parameters (*e.g.*, σ and s of Zimm and Bragg[3]), which were related to the intramolecular interactions. Among the various treatments, the nearest-neighbor interaction model (or the 2 × 2 matrix formulation) of Zimm and Bragg,[3] being an approximate version of their more detailed model (requiring an 8 × 8 statistical weight matrix[7,8]), has the simplest mathematical structure, and has been applied successfully to the analysis of numerous experimental data since it yields numerical values of σ and s which differ very little from those obtained with more elaborate theories involving larger matrices.

Recently it was shown that, even though only short-range interactions were assumed to be important in the original theories,[3–6] nevertheless the same mathematical formulation (*e.g.*, the 2 × 2 matrix treatment of Zimm and Bragg) still holds *effectively* for homopolymers, even when longer range intramolecular interactions are involved, provided that the empirical parameters σ and s are reinterpreted[9] so as to include contributions from these long-range interactions. However, as has been pointed out recently,[10] the 2 × 2 matrix formulation does not apply to specific-sequence copolymers of amino acids; *i.e.*, because of the different chemical nature of the individual residues, it is not possible to reinterpret σ and s (which vary for each type of residue in a copolymer) so as to include the effect of the longer range interactions, as was done for the homopolymer.[9] Therefore, in this paper, we develop a proper form for the statistical weight matrix for a copolymer, based on the approximation that the statistical weight assigned to the ith residue is independent of the amino acid type of its neighbors,[10] *i.e.*, side chain–backbone interactions (in a given solvent) determine the helix-forming power of each amino acid. Also, since the statistical weights in each matrix vary for each amino acid residue, it is not possible to carry out a similarity transformation (as is done with *homopolymers*) to simplify the evaluation of the partition function for a specific-sequence copolymer; hence, the matrix multiplication has to be carried out explicitly. It will be shown that we have to use an 11 × 11 matrix, which can be reduced further to a 4 × 4 matrix only for a *homopolymer*, by using a similarity transformation which depends on the amino acid type of the homopolymer. Fortunately, as will also be shown, the numerical results obtained for both homopolymers and copolymers with the 11 × 11 matrix are close (within a few per cent) to those obtained with the Zimm–Bragg 2 × 2 matrix. Therefore, unless the presently available precision in experimental data is improved, one can use the 2 × 2 matrix to analyze experimental data for a copolymer.

In section I, the model is described, and a mathematical formulation for the partition function is presented in section II. From this, the probability for each residue to be in one of eight possible states is calculated. In section III, we present and discuss the numerical results based on this model and compare them with those obtained with the 2 × 2 matrix formulation of Zimm and Bragg. Finally, in section IV, we apply this model to two proteins, myoglobin and lysozyme,

[1] This work was supported by research grants from the National Science Foundation (No. GB-28469X and GB-17388), from the National Institute of General Medical Sciences of the National Institutes of Health, U. S. Public Health Service (No. GM-14312), from the Eli Lilly, Hoffmann-La Roche, and Smith Kline and French Grants Committees, and from Walter and George Todd.

[2] (a) On leave of absence from the Department of Physics, Faculty of Science, University of Tokyo, Tokyo, Japan, 1967–1970 and summer, 1971; (b) National Research Council of Canada Postgraduate Fellow, 1971–1972.

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

[4] K. Nagai, *J. Phys. Soc. Jap.*, **15**, 407 (1960).

[5] A. Miyake, *J. Polym. Sci.*, **46**, 169 (1960).

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

[7] The 8 × 8 matrix has been shown previously,[8] in a heuristic fashion, to be contractable to a 4 × 4 matrix. A more systematic method for contracting statistical weight matrices is applied to this matrix in Appendix A.

[8] D. Poland and H. A. Scheraga, "Theory of Helix-Coil Transitions in Biopolymers," Academic Press, New York, N. Y., 1970, p 42.

[9] D. Poland and H. A. Scheraga, *Physiol. Chem. Phys.*, **1**, 389 (1969).

[10] M. Gō, N. Gō, and H. A. Scheraga, *J. Chem. Phys.*, **54**, 4489 (1971).

TABLE I
THE DEFINITION AND THE STATISTICAL WEIGHTS OF EIGHT STATES OF A RESIDUE

State	(ϕ, ψ)[a]	Hydrogen bond on[b]		Statistical weight	
		CO	NH	Real[c]	Dummy[d]
1	c	No	No	1	p_1
2	c	No	Yes	$\sigma^{-1/4}s^{1/2}$	p_2
3	c	Yes	No	$\sigma^{-1/4}s^{1/2}$	p_3
4	c	Yes	Yes	$\sigma^{-1/2}s$	p_4
5	h	No	No	$\sigma^{1/2}$	p_5
6	h	No	Yes	$\sigma^{1/4}s^{1/2}$	p_6
7	h	Yes	No	$\sigma^{1/4}s^{1/2}$	p_7
8	h	Yes	Yes	s	p_8

[a] The symbol h corresponds to the states of the dihedral angles ϕ and ψ of a residue in which they are confined to the small range in the (ϕ, ψ) space characteristic of the right-handed α helix; if the values of ϕ and ψ have no constraint at all, then the residue is in a c state. [b] The existence or absence of a hydrogen bond on the CO and NH groups is designated by yes or no, respectively. [c] The method of assignment of these eight statistical weights is discussed in ref 10. [d] These symbols are used in some of the formulas for the partition function, helix probabilities, etc. They assume the numerical values of the corresponding real statistical weights.

to illustrate how to define more precisely than heretofore[11] those portions of the denatured polypeptide chain which have a high propensity to take on an α-helical conformation. In Appendix A, a method for the systematic contraction of a statistical weight matrix is applied to the Zimm–Bragg 8 × 8 matrix, while the analogous copolymer statistical weight matrix for the Lifson–Roig model[6] is described in Appendix B; the Zimm–Bragg and Lifson–Roig treatments are also compared in Appendix B.

I. Description of the Model

In a recent paper[10] on the molecular theory of the helix–coil transition, a simple model (which is realistic to a first approximation) was proposed for the helix–coil transition in both homopolymers and copolymers of amino acids. This model was based on a knowledge of the important intramolecular interactions which are responsible for the helix–coil transition in polypeptides;[12] *i.e.*, cognizance was taken of the fact that an amino acid residue in a polypeptide chain can exist in any one of eight distinct states according to three *independent* factors: (a) whether or not the values of the two dihedral angles ϕ and ψ are constrained to those characteristic of the right-handed α helix, α_R, designated "h" and "c," respectively; (b) whether or not a hydrogen bond is formed between the CO group of the ith residue in question and the NH group of the $(i + 4)$ residue, when the three intermediate residues, (i + 1), $(i + 2)$, and $(i + 3)$, are all in h states (for the normal direction of the chain from the N to the C terminus); and (c) whether or not a hydrogen bond is formed between the NH group of the ith residue in question and the CO group of the $(i - 4)$ residue, when the three intermediate residues, $(i - 1)$, $(i - 2)$, and $(i - 3)$, are all in h states.

The definitions of the eight states are given in Table I, and an example of a conformational state of a polypeptide chain is given in Figure 1, in which each residue is designated as being in one of the eight possible states; the eight possible states of a residue are indicated by digits 1, 2, . . ., 8, and a conformation of the polypeptide chain is given by a linear sequence of these digits, one digit for each residue. A statistical weight, which depends uniquely on the type of amino acid residue and on the state of the residue, is assigned to each of the eight states (see Table I) and the statistical weight of a given conformational state of the chain is given by the product of the statistical weights, one for each residue. In the simplest model having the most essential features of the helix–coil transition in copolymers of amino acids, the parameter s of Table I is assumed to depend on the nature of the given amino acid but to be independent of the nature of its neighbors. In the numerical examples to be presented here, the parameter σ of Table I is assumed to be the same for all types of amino acids because reasonable variations in this parameter (*i.e.*, those found from experiments for different amino acids) have only a small effect on calculated average quantities. In practice, however, the model presented here allows for the use of a different value of σ for each amino acid type. The statistical weight depends only on the (type and) state of the residue in question, *i.e.*, it is independent of the (types and) states of the neighboring residues. However, this does not mean that the behavior of the system is noncooperative. The cooperativity arises through the correlation of the states that nearby residues can assume. For example, the existence of a hydrogen bond on the CO group of the ith residue (implying that the ith residue is in state 3, 4, 7, or 8) indicates that there is a hydrogen bond on the NH group of the $(i + 4)$ residue (*i.e.*, the $(i + 4)$ residue must be in state 2, 4, 6, or 8) because the hydrogen bond is formed between two amino acid residues separated by three intervening ones in h states. Conformations of a polypeptide chain consist of either all 1's (corresponding to the completely random-coil state with no α_R helical section) or of alternating coil and helical sections beginning and ending with a coil section (the dihedral angles of the end residues of the chain being unrestricted), a coil (respectively, a helical) section being defined as a string of residues that are all in the c (respectively h) states. The general forms of such alternating coil and helical sections are given in Table II. Any coil (respectively helical) section may follow any helical (respectively coil) section. A detailed discussion of the physical basis of this model is given in ref 10, and the development of the mathematical treatment of the model is given in section II. At the end of section II, the present model is compared with models for specific-sequence copolymers obtained by formally extending the Zimm–Bragg model for homopolymers.

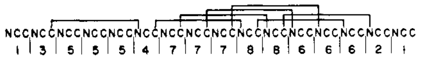

Figure 1. An example of a conformational state of a polypeptide chain consisting of 16 amino acid residues.

II. Mathematical Formulation

We now proceed to the development of the mathematical treatment of the model. Even though the latter was introduced primarily for the purpose of analyzing the helix–coil transition in specific-sequence copolymers of amino acids, it is instructive to apply it first to homopoly(amino acids), as a special case. For homopolymers, the method of

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

[12] Since the present theory is founded on a knowledge of the structure and the important intramolecular interactions, it is applicable only to the regular α helix, while the 2 × 2 matrix formulation of Zimm and Bragg for homopolymers is applicable even for such distorted helices as the 3_{10} helix or the α_{II} helix.[13]

[13] G. Nemethy, D. C. Phillips, S. J. Leach, and H. A. Scheraga, *Nature (London)*, **214**, 363 (1967).

Table II
General Form of Coil and Helical Sections[a]

First coil section of chain, followed by a helical section	Coil section flanked by helical sections
3	4
1 3	2 3
1 1 3	2 1 3
1 1 1 3	2 1 1 3
. . . . .	2 1 1 1 3
1 . . . 1 3	. . . . . .
	2 1 . . . 1 3
Last coil section of chain, following a helical section	Helical section flanked by coil sections
2	5 5 5
2 1	7 5 5 6
2 1 1	7 7 5 6 6
2 1 1 1	7 7 7 6 6 6
. . . . .	7 7 7 8 6 6 6
2 1 . . . 1	7 7 7 8 8 6 6 6
	7 7 7 8 8 8 6 6 6
	. . . .
	7 7 7 8 . . . 8 6 6 6

[a] Any conformation of a polypeptide, other than the one consisting of only 1's, is given by alternating coil and helical sections beginning and ending with a coil section. The "perfect" helix, *i.e.*, one with the maximum number of hydrogen bonds, has coil states at each end, and would be represented by the sequence 3 7 7 7 8 8 . . . 8 8 6 6 6 2.

sequence-generating functions[14] can be used. We confine ourselves to the discussion of an infinite chain, for which we can neglect end effects. Using the dummy statistical weights given in Table I, the sequence-generating functions $U(t)$ and $V(t)$ for an interior coil and helical section, respectively (see Table II) are given by

$$U(t) = \frac{p_4}{t} + \frac{p_2 p_3}{t^2}\left[1 - \frac{p_1}{t}\right]^{-1} \tag{1}$$

and

$$V(t) = \frac{p_5^3}{t^3} + \frac{p_5^2 p_6 p_7}{t^4} + \frac{p_5 p_6^2 p_7^2}{t^5} + \frac{p_6^3 p_7^3}{t^6}\left[1 - \frac{p_8}{t}\right]^{-1} \tag{2}$$

The secular equation is given by

$$U(t)V(t) = 1 \tag{3}$$

or, using eq 1 and 2

$$t^8 - (p_1 + p_8)t^7 + p_1 p_8 t^6 - p_4 p_5^3 t^4 - p_5^2\{p_4(p_6 p_7 - p_5 p_8) + p_5(p_2 p_3 - p_1 p_4)\}t^3 - p_5(p_6 p_7 - p_5 p_8)\{p_4 p_6 p_7 + p_5(p_2 p_3 - p_1 p_4)\}t^2 - p_6 p_7(p_6 p_7 - p_5 p_8)\{p_4 p_6 p_7 + p_5(p_2 p_3 - p_1 p_4)\}t - p_6^2 p_7^2(p_2 p_3 - p_1 p_4)(p_6 p_7 - p_5 p_8) = 0 \tag{4}$$

When the real values (fifth column of Table I) are substituted for the dummy statistical weights, the terms $(p_6 p_7 - p_5 p_8)$ and $(p_2 p_3 - p_1 p_4)$ vanish and eq 4 (after cancellation of a common factor of t^4) is reduced to

$$t^2(t - 1)(t - s) = \sigma s \tag{5}$$

which is identical with the secular equation obtained for the Zimm–Bragg model[3] (with $\mu = 3$), and also with that obtained for the Nagai model.[4] However, while the present model is identical with these two earlier models for homopolymers, it differs from them when applied to copolymers.

In order to apply the present model to specific-sequence copolymers, it is convenient to use the matrix method. From Table II, we can see that the possible state for the *i*th residue is determined by the states of three neighboring residues, taken here as the three preceding ones (*i.e.*, the $(i - 3)$, $(i - 2)$, and $(i - 1)$ residues). Therefore, the rows of the statistical weight matrix correspond to the states of residues $(i - 3)$, $(i - 2)$, and $(i - 1)$ and the columns to the states of residues $(i - 2)$, $(i - 1)$, and (i), and the elements of the matrix will have two three-digit suffixes, the first corresponding to the triplet defined by the row indices and the second to the triplet defined by the column indices. However, according to the formulation of the model (Table I), of the 8^3 possible triplets, only 44 are physically real ones (*e.g.*, all triplets in which state 5 follows state 2 cannot occur); for the same reason, only 72 quadruplets have physical reality. The 44 possible triplets and the 72 possible quadruplets are listed in Table III. Since there are 44 possible triplets, the size of the statistical weight matrix is 44 × 44. This matrix **V** is given in Table IV. Two kinds of zeros occur as elements in this matrix. The first (shown by a blank in Table IV) is a nonsense zero when the two suffixes are inconsistent with each other [*e.g.*, element (355, 756)]; the second (shown positively by a 0 in Table IV to distinguish it from a nonsense zero) arises when two allowable triplets lead to a disallowed quadruplet, *i.e.*, when the quadruplet is not one of the 72 possible ones [*e.g.*, element (355, 556)]. When two allowable triplets do define an allowed quadruplet [one of the 72 possible ones, *e.g.*, element (355, 555)], then the value of the element is nonvanishing. Therefore, there are only 72 nonvanishing elements out of the 44 × 44 elements in the matrix. The value of the nonvanishing element is determined only by the state of the *i*th residue, since the correlation with respect to residues $(i - 3)$, $(i - 2)$, and $(i - 1)$ is built into the designation of the states by the numbers 1–8; thus, for example, if the *i*th residue is in state 3, the statistical weight assigned is p_3, irrespective of the states of residues $(i - 3)$, $(i - 2)$, and $(i - 1)$.

Table III
All Possible Triplets and Quadruplets (Zimm–Bragg Model)

$i-3$	$i-2$	$i-1$	i	$i-3$	$i-2$	$i-1$	i
1	1	1	1, 3	3	7	5	5
2	1	1	1, 3	4	7	5	5
5	2	1	1, 3	7	7	5	6
6	2	1	1, 3	5	5	6	2, 4
5	5	2	1, 3	7	5	6	6
5	6	2	1, 3	5	6	6	2, 4
6	6	2	1, 3	6	6	6	2, 4
1	1	3	5, 7	7	6	6	6
2	1	3	5, 7	8	6	6	6
5	2	3	5, 7	7	7	6	6
6	2	3	5, 7	7	8	6	6
5	5	4	5, 7	8	8	6	6
5	6	4	5, 7	1	3	7	5, 7
6	6	4	5, 7	2	3	7	5, 7
1	3	5	5	5	4	7	5, 7
2	3	5	5	6	4	7	5, 7
5	4	5	5	3	7	7	5, 7
6	4	5	5	4	7	7	5, 7
3	5	5	5	7	7	7	6, 8
4	5	5	5	7	7	8	6, 8
5	5	5	2, 4	7	8	8	6, 8
7	5	5	6	8	8	8	6, 8

The partition function Z of a specific-sequence copolymer of N amino acids, with the jth one being of amino acid type $A(j)$, is given by eq 6

[14] S. Lifson, *J. Chem. Phys.*, **40**, 3705 (1964).

TABLE IV
44 × 44 MATRIX V

				α							β							γ				δ		ε		ζ				η				θ				ι				κ		λ			
			i	1	1	1	1	2	2	2	3	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	6	6	6	5	6	6	6	5	6	6	6	7	7	7	7	7	7	7	8	8	8
			$i-1$	1	1	2	2	5	6	6	1	1	2	2	5	6	6	3	3	4	4	5	5	7	7	5	5	6	6	5	5	6	6	7	7	8	8	3	3	4	4	7	7	7	7	8	8
	$i-3$	$i-2$	$i-1$ / $i-2$	1	2	5	6	5	5	6	1	2	5	6	5	5	6	1	2	5	6	3	4	3	4	5	5	5	6	7	7	7	8	7	7	7	8	1	2	5	6	3	4	7	7	7	8
α	1	1	1	p_1							p_3																																				
	2	1	1	p_1							p_3																																				
	5	2	1		p_1							p_3																																			
	6	2	1		p_1							p_3																																			
	5	5	2			p_1							p_3																																		
	5	6	2				p_1							p_3																																	
	6	6	2				p_1							p_3																																	
β	1	1	3															p_5																				p_7									
	2	1	3															p_5																				p_7									
	5	2	3																p_5																				p_7								
	6	2	3																p_5																				p_7								
	5	5	4																	p_5																				p_7							
	5	6	4																		p_5																				p_7						
	6	6	4																		p_5																				p_7						
γ	1	3	5																			p_5																									
	2	3	5																			p_5																									
	5	4	5																				p_5																								
	6	4	5																				p_5																								
δ	3	5	5					0							0											p_5	0																				
	4	5	5					0							0											p_5	0																				
ε	3	7	5																											p_5	0																
	4	7	5																											p_5	0																
ζ	5	5	5					p_2							p_4											0	0																				
	5	5	6						p_2							p_4												0																			
	5	6	6							p_2							p_4												0																		
	6	6	6							p_2							p_4												0																		
η	7	5	5					0							0											0	p_6	p_6																			
	7	5	6						0							0																															
	7	6	6							0							0												p_6																		
	8	6	6							0							0												p_6																		
θ	7	7	5																											0	p_6																
	7	7	6																													p_6															
	7	8	6																														p_6														
	8	8	6																														p_6														
ι	1	3	7																					p_5																		p_7					
	2	3	7																					p_5																		p_7					
	5	4	7																						p_5																		p_7				
	6	4	7																						p_5																		p_7				
κ	3	7	7																															p_5	0									p_7	0		
	4	7	7																															p_5	0									p_7	0		
λ	7	7	7																															0	p_6									0	p_8		
	7	7	8																																	p_6										p_8	
	7	8	8																																		p_6										p_8
	8	8	8																																		p_6										p_8

Table V

Contracted 11 × 11 Matrix W

	α	β	γ	δ	ϵ	ζ	η	θ	ι	κ	λ
α	p_1	p_3	0	0	0	0	0	0	0	0	0
β	0	0	p_5	0	0	0	0	0	p_7	0	0
γ	0	0	0	p_5	0	0	0	0	0	0	0
δ	0	0	0	0	0	p_5	0	0	0	0	0
ϵ	0	0	0	0	0	0	p_5	0	0	0	0
ζ	p_2	p_4	0	0	0	0	0	0	0	0	0
η	0	0	0	0	0	p_6	0	0	0	0	0
θ	0	0	0	0	0	0	p_6	0	0	0	0
ι	0	0	0	0	p_5	0	0	0	0	p_7	0
κ	0	0	0	0	0	0	0	p_5	0	0	p_7
λ	0	0	0	0	0	0	0	p_6	0	0	p_8

$$Z = \mathbf{s} \prod_{j=1}^{N} \mathbf{V}_{\mathrm{A}(j)} \mathbf{t} \tag{6}$$

where $\mathbf{V}_{\mathrm{A}(j)}$ is the 44 × 44 matrix of Table IV with the values of the elements corresponding to amino acid type A(*j*), and **s** and **t** are the following row and column vectors, respectively, with 44 elements each.

$$\mathbf{s} = (1, 0, 0, 0, 0, 0, 0, 0, 0, \ldots, 0) \tag{7}$$

and

$$\mathbf{t} = (1, 1, 1, 1, 1, 1, 1, 0, 0, \ldots, 0)^{+} \tag{8}$$

where the plus sign designates a transpose. The end vectors **s** and **t** are determined from the general form of the terminal coil sections shown in Table II. From Table IV, it can be seen that the **s** vector corresponds to the situation in which the *i*th residue, as the first one in the chain, is in state 1 or 3 and is preceded by three imaginary residues in state 1, *i.e.*, to the quadruplet states 1111 and 1113. Similarly, the **t** vector corresponds to the situation in which the *i*th residue, as the last one in the chain, is in state 1 or 2 and is preceded by any doublet states which enable the last triplet in the chain to be any one of the following allowable ones, *viz.*, 111, 211, 521, 621, 552, 562, 662.

The first and second rows of the 44 × 44 matrix given in Table IV are identical Similarly, there are many other pairs of identical rows. This implies that the 44 × 44 matrix **V** can be contracted. In fact, it can be contracted to an 11 × 11 matrix **W**, as shown below. In order to perform the contraction, the 44 possible triplets are classified into 11 groups designated as α through λ, and the 44 × 44 matrix **V** is divided into submatrices according to this classification of the possible triplets, as shown in Table IV. Thus, the matrix **V** consists of 11^2 or 121 submatrices. Each submatrix is either of the following two forms: (a) all elements are vanishing or (b) all rows have one and only one nonvanishing element, the values of all the nonvanishing elements of a particular submatrix being identical. The contracted 11 × 11 matrix **W** is obtained by replacing each submatrix of **V** by a number, zero for type-a submatrices and the common value of the nonvanishing element for each type-b submatrix. Thus, each of the two three-digit siffixes of the elements of the matrix **W** may be replaced by the indices α through λ. The contracted matrix **W** is given in Table V. Just as the matrix **V** was divided into submatrices, similarly each of the 44-dimensional vectors **s** and **t** can also be divided into 11 subvectors by grouping their components according to the classification of the possible triplets. Thus, the subvector of the row vector **s**, designated as α, is a seven-dimensional row vector whose first component is unity, with all other six components vanishing; all subvectors of **s** other than α are zero vectors, *i.e.*, all components vanish. Hence, **s** is contracted to the 11-dimensional vector

$$\mathbf{u} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \tag{9}$$

by replacing each subvector by a number, unity for nonzero subvectors and zero for zero subvectors. A subvector of the column vector **t**, designated as α, is a seven-dimensional column vector whose components are all unity; all other subvectors of **t** are zero vectors. Hence, **t** is contracted to the 11-dimensional vector

$$\mathbf{v} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^{+} \tag{10}$$

The partition function of the system is then given by

$$Z = \mathbf{u} \prod_{j=1}^{N} \mathbf{W}_{\mathrm{A}(j)} \mathbf{v} \tag{11}$$

where eq 11 is derived from eq 6 by direct application of the rules for matrix multiplication, as follows. The right-hand side of eq 6 can be expressed as a sum of products of the subvectors and submatrices mentioned above. It is to be noted here that a product of type-b submatrices is also a type-b submatrix with the value of the common nonvanishing elements being equal to the product of the common nonvanishing elements of each submatrix. The operation of multiplying the nonzero row subvector of **s**, designated as α, and the nonzero column subvector of **t**, also designated as α, from left and right, respectively, is equivalent to picking up a number having the value of the product of the common nonvanishing elements of each nonzero submatrix. Thus, each product of subvectors and submatrices in the expanded form of eq 6 is simply equal to a product of the common nonvanishing elements of each submatrix. Therefore, the partition function of eq 6, which is a sum of such products, can be expressed in another way (*viz.*, as eq 11) by replacing the nonzero subvectors and submatrices by a number having the value of their common nonvanishing elements. While this proof demonstrates how eq 11 was derived from eq 6, the identity of these two equations is also indicated below by showing that the contraction is equivalent to a similarity transformation.

Since the secular equation for the matrix **W** is the same as that obtained from the method of sequence-generating functions (the latter applying, of course, only to a homopolymer) we may express the secular equation for the 11 × 11 matrix **W** in terms of *U*(*t*) and *V*(*t*) given by eq 1 and 2, respectively; it is found to have the form

$$t^{8}\{U(t)V(t) - 1\} = 0 \tag{12}$$

or

$$t^{7}\{t^{2}(t - 1)(t - s) - \sigma s\} = 0 \tag{13}$$

where the quantity in braces in eq 13 is the fourth-degree polynomial of eq 5. From the form of eq 13, the factor t^7 can be dropped, and the secular equation of **W** is then of the fourth degree, *i.e.*, the rank of the 11 × 11 matrix **W** is four. The question then arises as to whether the 11 × 11 matrix **W** can be contracted further to a 4 × 4 matrix. In order to clarify this point, eq 11 is derived below by a different method, which enables us to interpret the contraction (from

44×44 to 11×11) as a similarity transformation of vectors and matrices.

As was already noted, there are many pairs of identical rows in the 44×44 matrix $\mathbf{V}$. Thus, let us assume that the ith and jth rows of $\mathbf{V}$ are identical. Then we shall perform two successive similarity transformations in which $\mathbf{s}$, $\mathbf{V}$, and $\mathbf{t}$ are transformed to $\mathbf{s}'$, $\mathbf{V}'$, and $\mathbf{t}'$, respectively, *viz.*

$$\begin{aligned} \mathbf{s}' &= \mathbf{s}\mathbf{U}_1^{-1}\mathbf{U}_2^{-1} \\ \mathbf{V}' &= \mathbf{U}_2\mathbf{U}_1\mathbf{V}\mathbf{U}_1^{-1}\mathbf{U}_2^{-1} \\ \mathbf{t}' &= \mathbf{U}_2\mathbf{U}_1\mathbf{t} \end{aligned} \tag{14}$$

i.e.,

$$\mathbf{s}'\mathbf{V}'\mathbf{t}' = \mathbf{s}\mathbf{V}\mathbf{t} \tag{15}$$

By generalization of eq 14 and 15, it follows that the partition function, eq 6, is invariant under these transformations. The matrices $\mathbf{U}_1$, $\mathbf{U}_1^{-1}$, $\mathbf{U}_2$, and $\mathbf{U}_2^{-1}$ are chosen as

$$\mathbf{U}_1 = \begin{array}{c|cccccccc|cccc} & & & & & & & & j & & & & \\ \hline & 1 & & & & & & & & & & & \\ & & \cdot & & & & & & & & & & \\ & & & \cdot & & & & & & & & & \\ & & & & \cdot & & & & & & & & \\ i & & & & & 1 & & & -1 & & & & \\ \hline & & & & & & 1 & & & & & & \\ & & & & & & & \cdot & & & & & \\ & & & & & & & & \cdot & & & & \\ & & & & & & & & \cdot & & & & \\ & & & & & & & & 1 & & & & \\ & & & & & & & & & 1 & & & \\ & & & & & & & & & & \cdot & & \\ & & & & & & & & & & & \cdot & \\ & & & & & & & & & & & & 1 \end{array} \tag{16}$$

$$\mathbf{U}_1^{-1} = \begin{array}{c|cccccccc|cccc} & & & & & & & & j & & & & \\ \hline & 1 & & & & & & & & & & & \\ & & \cdot & & & & & & & & & & \\ & & & \cdot & & & & & & & & & \\ & & & & \cdot & & & & & & & & \\ i & & & & & 1 & & & 1 & & & & \\ \hline & & & & & & 1 & & & & & & \\ & & & & & & & \cdot & & & & & \\ & & & & & & & & \cdot & & & & \\ & & & & & & & & \cdot & & & & \\ & & & & & & & & 1 & & & & \\ & & & & & & & & & 1 & & & \\ & & & & & & & & & & \cdot & & \\ & & & & & & & & & & & \cdot & \\ & & & & & & & & & & & & 1 \end{array} \tag{17}$$

$$\mathbf{U}_2 = \begin{array}{c|ccccc|cccc} & & & & & i & & & & \\ \hline & 0 & & & & 1 & & & & \\ & 1 & \cdot & & & & & & & \\ & & \cdot & \cdot & & & & & & \\ & & & \cdot & \cdot & & & & & \\ i & & & & 1 & 0 & & & & \\ \hline & & & & & & 1 & & & \\ & & & & & & & \cdot & & \\ & & & & & & & & \cdot & \\ & & & & & & & & & 1 \end{array} \tag{18}$$

$$\mathbf{U}_2^{-1} = \begin{array}{c|ccccc|cccc} & & & & & i & & & & \\ \hline & 0 & 1 & & & & & & & \\ & & \cdot & \cdot & & & & & & \\ & & & \cdot & \cdot & & & & & \\ & & & & \cdot & 1 & & & & \\ i & 1 & & & & 0 & & & & \\ \hline & & & & & & 1 & & & \\ & & & & & & & \cdot & & \\ & & & & & & & & \cdot & \\ & & & & & & & & & 1 \end{array} \tag{19}$$

Since the ith and jth rows of $\mathbf{V}$ are identical, $\mathbf{V}'$ will have the following form, for $\mathbf{U}_1$, $\mathbf{U}_1^{-1}$, $\mathbf{U}_2$, $\mathbf{U}^{-1}$ given by eq 16–19

$$\mathbf{V}' = \begin{bmatrix} 0 & \mathbf{0} \\ \mathbf{a} & \mathbf{V}'' \end{bmatrix} \tag{20}$$

where $\mathbf{0}$ is a 43-dimensional zero row vector, $\mathbf{a}$ is a certain 43-dimensional (nonzero) column vector (whose magnitude is of no concern to us), and $\mathbf{V}''$ is a 43×43 matrix. According to eq 14, the vectors $\mathbf{s}'$ and $\mathbf{t}'$ are of the form

$$\mathbf{s}' = (b, \mathbf{s}'') \tag{21}$$

$$\mathbf{t}' = (0, \mathbf{t}'')^{+} \tag{22}$$

where $\mathbf{s}''$ and $\mathbf{t}''$ are 43-dimensional row and column vectors whose ith components are the same as the $(i + 1)$ components of $\mathbf{s}'$ and $\mathbf{t}'$, respectively, and b is a number which does not appear in the final expression. The first component of $\mathbf{t}'$ is zero, irrespective of which pair of identical rows of $\mathbf{V}$ is chosen for performance of the transformation.

If the second of eq 14 is used N times, then the partition function of eq 6 becomes

$$Z = \mathbf{s}' \prod_{j=1}^{N} \mathbf{V}'_{\mathrm{A}(j)}\mathbf{t}' \tag{23}$$

which is identical with eq 6. Equation 23 may be written, by substitution of eq 20–22, as

$$Z = (\mathrm{b}, \mathbf{s}'')\begin{bmatrix} 0 & \mathbf{0} \\ \left\{\prod_{j=1}^{N-1} \mathbf{V}''_{\mathrm{A}(j)}\right\}\mathbf{a}_{\mathrm{A}(N)} & \prod_{j=1}^{N} \mathbf{V}''_{\mathrm{A}(j)} \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{t}'' \end{bmatrix} \tag{24}$$

or as

$$Z = \mathbf{s}'' \prod_{j=1}^{N} \mathbf{V}''_{\mathrm{A}(j)} \mathbf{t}'' \qquad (25)$$

This equation indicates that the 44 × 44 matrix has been contracted to a 43 × 43 matrix. It is possible to carry out the same procedure again as long as (a) there are one or more pairs of identical rows in the contracted matrix, and (b) the first element of the transformed column vector $\mathbf{t}'$ is zero. By applying the series of contraction procedures mentioned above to eq 6 (involving the 44 × 44 matrix of Table IV), we arrive at eq 11, which involves the 11 × 11 matrix of Table V. From the above procedure, we see that the contraction of the statistical weight matrix can be interpreted as a similarity transformation.

Since there is no pair of identical rows in the **W** matrix of Table V, it cannot be contracted any further by the above procedure. Now, as already shown (see eq 13), the rank of the **W** matrix is four. For a homopolymer, there is a similarity transformation which contracts the 11 × 11 **W** matrix into a 4 × 4 matrix. However, it turns out empirically that the elements of the matrix required for any such similarity transformation are functions of the statistical weights, p_i, instead of being 1's or 0's; hence, such a similarity transformation cannot be used to simplify eq 12 for copolymers, because a matrix **U** of a similarity transformation which leads to contraction of the **W** matrix for one type of amino acid does not contract the **W** matrix for other types of amino acid. This is the reason why the 11 × 11 matrix with rank 4 must be used for the analysis of the helix-coil transition in copolymers.

The probability $P_i(j)$ that the jth residue is in state i, with statistical weight p_i, is given by

$$P_i(j) = \mathbf{u}\left[\prod_{k=1}^{j-1} \mathbf{W}_{\mathrm{A}(k)}\right] \frac{\partial \mathbf{W}_{\mathrm{A}(j)}}{\partial \ln p_i(j)} \left[\prod_{k=j+1}^{N} \mathbf{W}_{\mathrm{A}(k)}\right] \mathbf{v}/Z \qquad (26)$$

This is an extension of eq 2 of ref 11, which was developed for the calculation of the helix probability profiles of denatured proteins.

At this point, it will be very useful to compare the present model with those for specific-sequence copolymers which can be obtained by formally extending the Zimm–Bragg model for homopolymers. In the Zimm–Bragg model[3] *for a homopolymer* with $\mu = 3$, the two allowed states for each amino acid residue are designated 1 or 0, depending only on whether there is or is not, respectively, a hydrogen bond between the NH group of the ith residue in question and the CO group of the $(i - 4)$ residue[15] (for the normal direction of the chain from the N to the C terminus). From this definition, it follows that an isolated 0 or an isolated pair of 0's cannot occur, and hence that it is necessary to know the states of four residues, say, $i - 3$, $i - 2$, $i - 1$, and i, in order to assign a statistical weight to the ith residue. The rule for the assignment of statistical weights in the Zimm–Bragg model with $\mu = 3$ is as follows: 1 is assigned to a 0 state, s to a 1 state following a 1 state, and σs to a 1 state following three or more 0 states. Such a correlation would require an 8 × 8 statistical weight matrix[3,17] to begin with, but this 8 × 8 matrix can be contracted to a 4 × 4 matrix, as shown in Appendix A.

We can formally extend the Zimm–Bragg model for a homopolymer with $\mu = 3$ to a copolymer of N amino acids if we use different values of s (assuming σ to be the same) for different amino acids. For such a model, the partition function (see Appendix A) is given by

$$Z = (1, 0, 0, 0) \prod_{j=5}^{N} \mathbf{W}_{\mathrm{A}(j)} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad (27)$$

where $\mathbf{W}_{\mathrm{A}(j)}$ is given by

$$\mathbf{W}_{\mathrm{A}(j)} = \begin{bmatrix} 1 & 0 & 0 & \sigma s_{\mathrm{A}(j)} \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & s_{\mathrm{A}(j)} \end{bmatrix} \qquad (28)$$

The product in eq 27 starts at $j = 5$ because the first four NH groups of the chain are always unbonded. Equation 27 is simpler than eq 11, because the matrix size is smaller. The statistical weight of a helical sequence consisting of l hydrogen bonds, in this extension of the Zimm–Bragg model, is

$$\sigma \prod_{j=k}^{k+l-1} s_{\mathrm{A}(j)}$$

For the model described in section I, the corresponding statistical weight (from Table I) is

$$\sigma[s_{\mathrm{A}(k)} s_{\mathrm{A}(k+1)} s_{\mathrm{A}(k+2)} s_{\mathrm{A}(k+3)}]^{1/2} \left[\prod_{i=k+4}^{k+l-1} s_{\mathrm{A}(i)}\right] \times [s_{\mathrm{A}(k+l)} s_{\mathrm{A}(k+l+1)} s_{\mathrm{A}(k+l+2)} s_{\mathrm{A}(k+l+3)}]^{1/2}$$

In general, the values of these two expressions will be different for a copolymer because the value of $s_{\mathrm{A}(j)}$ varies with the amino acid type. However, if eq 27 were used for a copolymer, it would imply the validity of an assumption about the intramolecular interactions, which is physically unrealistic, as is shown by the following argument. The value of the parameter s for a particular amino acid residue is assumed to depend only on the type of the amino acid residue in question, and *not* on the types of neighboring amino acid residues (this same assumption is made in the new model described in section I, and is the basis for the assignments of the statistical weights shown in Table I). This assumption means that side chain–side chain interactions are negligible; if such interactions were not negligible, the statistical weight of a residue would depend on the type of neighboring residues. This assumption has been found to be valid for polyglycine and poly-L-alanine by an analysis[10] of the intramolecular interactions responsible for the stability of an α helix, wherein it was shown that *intraresidue* interactions (and a hydrogen-bond energy, which is independent of amino acid type) dominate all others; *i.e.*, the assumed dependence of the parameter s on the type of amino acid residue arises from the interaction between the side chain of the residue in question and the backbone of the peptide chain. Now, if the Zimm–Bragg model with $\mu = 3$ (in contrast to *our* model) is extended to copolymers, the statistical weight of a residue i is determined

[15] In the original paper of Zimm and Bragg,[3] an NH–CO group, rather than an amino acid residue (as used here), was treated as the unit structure. However, if we follow the new convention for the nomenclature of polypeptide chains,[16] the Cα–CO–NH group should be taken as the unit, and therefore the CO of the ith unit is hydrogen bonded to the NH of the $(i + 3)$ unit in the Zimm–Bragg model. In terms of residues, this would mean that the CO of the ith residue is hydrogen bonded to the NH of the $(i + 4)$ residue. However, we have changed the definition of a 1 state, as given in the text, to maintain the Zimm–Bragg rules for the statistical weights; this alteration in the definition of the 1 state compensates for the fact that the Zimm–Bragg definition does not conform to the new convention.

[16] IUPAC–IUB Commission on Biochemical Nomenclature, *Biochemistry*, 9, 3471 (1970).

[17] Reference 8, p 34.

TABLE VI
SEQUENCE OF THE RANDOM COPOLYMER BETWEEN RESIDUES 400 AND 700, CORRESPONDING TO THE ABSCISSA IN FIGURES 2 AND 3[a]

400 410 420 430 440
ABBABAABBABABBBBBABABAAABAAABBAABBBBBBABABABAB*AAAA*
450
*A*BAB*AAAA*BBBABBBBBAAABBABBAAABBBBAABABBAABBABBABBAA
500
BAABABAABBBBABBBBBBBBABABAAABABBBBAAABABBBAABBBBAB
550
ABBBABBAB*AAAA*BBBBB*AAAA*BABABABABBAABBBBBBBBBBB*AAAA*B
600
BABBABABBABAAB*AAAAAA*BAABBAABBAABABABBBABABABBBBAAB
650 700
BBBAABAAAB*AAAAAAA*BBBBB*AAAAA*BBBABBBBBBBBAAABAB*AAAAAA*

[a] Regions of four or more consecutive A's (strong helix formers) are in italics.

by the type and state of the residue in question and by the states of the three preceding residues (correlation of states of residues $i - 3$, $i - 2$, $i - 1$, and i). The states of residues $i + 1$, $i + 2$, etc., do not influence the statistical weight of the residue i. This procedure would be applicable to a homopolymer but not to a copolymer because it implies that the conformations of the (different) types of residues $i + 1$, $i + 2$, etc. (in a copolymer) do not affect the conformation of the ith residue, an implication which is quite unrealistic. In the new model described in section I, the statistical weight of a residue is determined only by the type and state of the ith residue; however, the state of the ith residue is correlated with the states of residues $i - 3$, $i - 2$, $i - 1$ *and* $i + 1$, $i + 2$, $i + 3$. In other words, the copolymer character, which appears properly in eq 11 but *not* in eq 27, arises from the fact that s is different for each different type of amino acid and the cooperative character from the correlation of the states of neighboring residues.

In the case of homopolymers, the 2 × 2 matrix formulation of Zimm and Bragg[3] (the Zimm–Bragg model with $\mu = 1$) is a very good approximation to the Zimm–Bragg model with $\mu = 3$. Even though such a simple approximation (*i.e.*, use of a 2 × 2 matrix) is not applicable to copolymers, nevertheless it has been used[11] for copolymers (and its use will be justified in sections III and IV) because of its mathematical simplicity. In such an application of the 2 × 2 matrix treatment to a copolymer of N amino acid residues, the partition function is given by

$$Z = (1, 0)\left[\prod_{j=1}^{N} \mathbf{W}_{\mathrm{A}(j)}\right]\begin{bmatrix}1\\1\end{bmatrix} \tag{29}$$

Here, $\mathbf{W}_{\mathrm{A}(j)}$ is given by

$$\mathbf{W}_{\mathrm{A}(j)} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix} \tag{30}$$

where the real values of the dummy variables p_1–p_4 (*not* the same as those of Table I) are given by

$$\begin{aligned} p_1 &= 1 \\ p_2 &= \sigma s_{\mathrm{A}(j)} \\ p_3 &= 1 \\ p_4 &= s_{\mathrm{A}(j)} \end{aligned} \tag{31}$$

In this treatment, each amino acid is considered to be in one of two states, h or c, or helical and coil. However, the specification of a residue as helical is arbitrary to some extent. For example, a residue can be specified as helical by using one or more of the three independent factors discussed in section I. The state h may be interpreted as a helical one in the sense that it conforms to any one of the three factors. By using a formula similar to eq 26, in which differentiation with respect to the dummy variables appearing in eq 30 and subsequent substitution of the real values of eq 31 are involved, one can calculate the probabilities that the jth residue is helical and follows a residue in a helical or a coil state, $P_{\mathrm{hh}}(j)$ and $P_{\mathrm{ch}}(j)$ respectively, and that it is in a coil state and follows a residue in a helical or a coil state, $P_{\mathrm{hc}}(j)$ and $P_{\mathrm{cc}}(j)$, respectively. The quantity $p_{\mathrm{H}}(j)$, the probability that a residue is in a helical state, is simply given by the sum of $P_{\mathrm{hh}}(j)$ and $P_{\mathrm{ch}}(j)$. The quantities P_{ch} and P_{hc} may be interpreted as the probabilities that a residue is the first one in a helical or a coil section, respectively. However, because of the arbitrariness in the definition of a helical residue, the phrase "first residue of a helical or a coil section" is not defined precisely. From Table II, one can see that the junctions between long coil and helical sections have the following structures: *i.e.*, 137778 for a junction at which a helical section follows a coil section, and 866621 for a junction at which a coil section follows a helical section. For this case, the 2 × 2 matrix formulation does not specify precisely which one of the residues at these junctions is the first residue. Therefore, the new formulation developed in this paper, which is able to distinguish eight states of a residue in Table I, contains much more information than the 2 × 2 matrix formulation. It should also be noted that, in the 2 × 2 matrix formulation, the probabilities that a residue is in states 4 and 5 cannot be calculated, even by calculating $P_{\mathrm{ch}}(j)$ and $P_{\mathrm{hc}}(j)$. In spite of the arbitrariness discussed above, it is interesting to see how the results of the 2 × 2 matrix formulation compare with the results obtained by using the new formulation. This comparison is made in sections III and IV.

The model described in section I, and formulated mathematically in this section, is essentially that of Zimm and Bragg[3] applied to specific-sequence copolymers. The analogous model for the Lifson–Roig[6] formulation of the helix–coil transition is given in Appendix B. In addition, the differences between the Zimm–Bragg and Lifson–Roig models are considered, so that explicit expressions for the conversion of parameters from one model to those of the other can be derived. It is shown that, as long as the parameters are correctly related, the models are essentially equivalent.

III. Numerical Examples

The formulation developed in section II is applied here to the calculation of the conformation of a specific-sequence (random) copolymer of two types of amino acids A and B The chain length was taken as 1000, and the sequence of A's and B's was obtained numerically with a random number

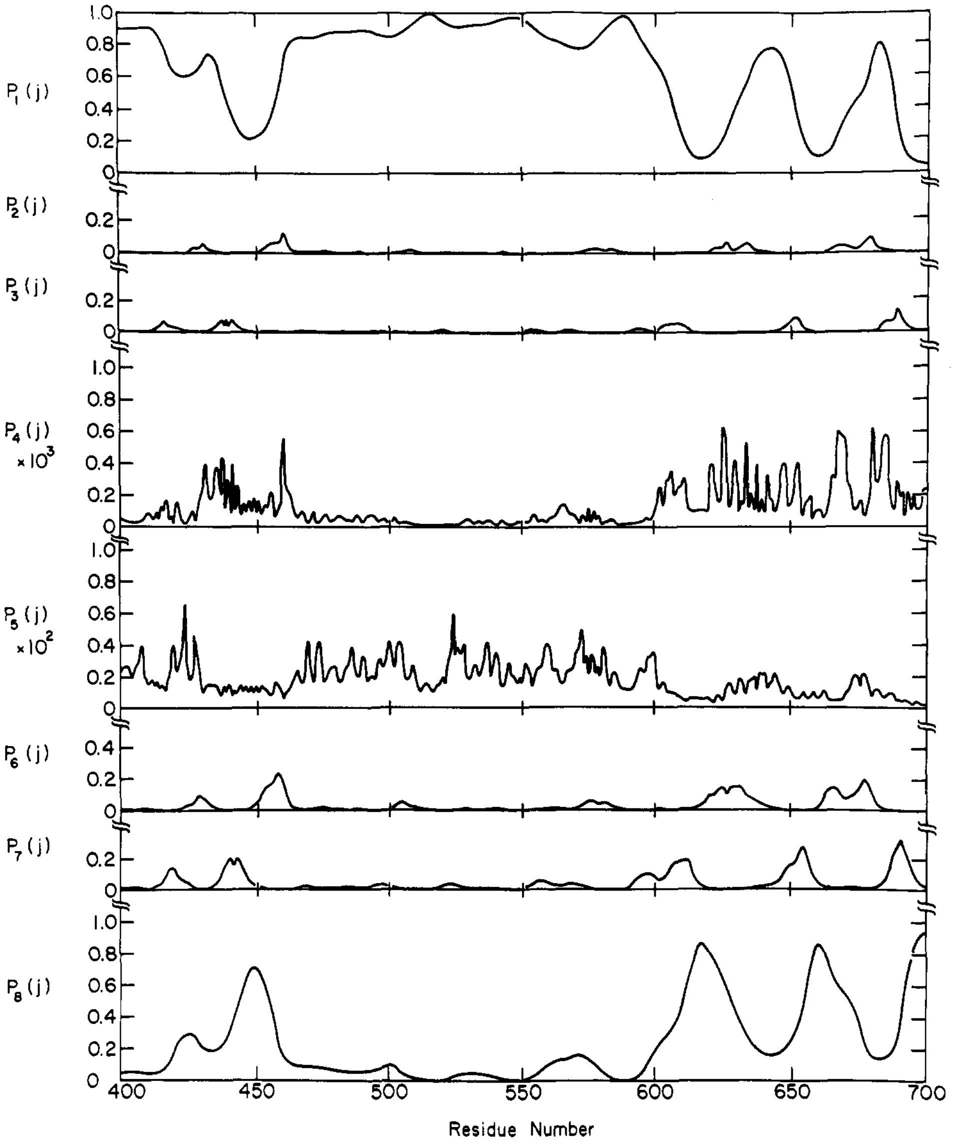

Figure 2. Curves of $P_i(j)$ for $i = 1, 2, 3, \ldots, 8$ and j = 400–700, *i.e.*, for residues in the central portion of the random copolymer shown in Table VI. $s_A = 2.01$, $s_B = 0.50$, and $\sigma = 5 \times 10^{-4}$; hence, these curves pertain to a fixed temperature. With the exception of P_4 and P_5, the ordinate has the units of normalized probability and the abscissa has units of chain site j. The probability is raised by a factor of 10^3 and 10^2 for P_4 and P_5, respectively.

generator;[18] the symbols A and B were assigned to numbers less than or greater than 0.5, respectively (*i.e.*, the fraction of A and B units in the copolymer was taken as 0.5). A portion of the amino acid sequence between residues 400 and 700 is shown in Table VI. The two types of amino acids were characterized by the values of their s parameters, s_A and s_B, respectively (see captions of Figures 2 and 3). The value of the nucleation parameter σ, assumed to be the same for A and B, was taken as 5×10^{-4}. The helix probabilities for this copolymer were calculated both by eq 26 and also by use of the 2×2 matrix formulation, and the two sets of results are compared below.

In Figure 2, calculated probabilities $P_i(j)$, for the jth residue to be in state i, are plotted for $s_A = 2.01$ and $s_B = 0.50$, for $i = 1, 2, \ldots, 8$ and j = 400–700, *i.e.*, for residues in the central portion of the specific-sequence (random) copolymer shown in Table VI. The probabilities for each residue to be in state 3, 4, 7, or 8 [designated as $P_{3,4,7,8}(j)$] and in state 5, 6, 7, or 8 [designated as $P_{5,6,7,8}(j)$], as well as the probability for each residue to be in the "helical state" in the sense of the 2×2 matrix formulation [designated as $P_H(j)$], are plotted in Figure 3. In the original treatment by Zimm and Bragg, each residue can be in either a "helical" or a "coil" state, a helical state being defined as one in which the CO group of the residue is involved in the formation of an intrachain hydrogen bond. In the formulation developed in this paper, the CO group of a residue is involved in the formation of an intrachain hydrogen bond in states 3, 4, 7, and 8 (see Table I). Therefore, if we interpret $P_H(j)$ as the probability of the jth residue to be helical in the original sense of Zimm and Bragg, $P_H(j)$ must be compared directly with $P_3(j) + P_4(j) + P_7(j) + P_8(j)$, designated as $P_{3,4,7,8}(j)$. Another possible definition of a residue in a helical state would be that in which the residue is in the h state defined in Table I (*i.e.*, dihedral angles ϕ and ψ of the residue are confined to a small range of the (ϕ, ψ) space characteristic of the right-handed α helix designated as α_R). Because a residue is in the h state in states 5, 6, 7, and 8, the probability for the jth residue to be helical in this sense is given by $P_5(j) + P_6(j) + P_7(j) + P_8(j)$, designated as

[18] With the IBM 360/65 computer and an IBM scientific subroutine package, subroutine RANDU, where IX, a user-supplied parameter, was taken as 999,999.

TABLE VII
THREE DIFFERENT HELIX CONTENTS IN PER CENT FOR A SPECIFIC-SEQUENCE COPOLYMER OF TWO KINDS OF UNIT[a]

$s_A = 0.50$	$s_B = 1.00$	1.22	1.49	1.82	2.23	2.72	3.32	4.06	4.95
θ_H	0.5	1.3	5.7	25.0	52.3	74.4	84.9	89.6	92.0
$\theta_{3,4,7,8}$	0.5	1.1	4.5	22.4	51.5	76.8	88.2	92.8	95.1
$\theta_{5,6,7,8}$	0.7	1.5	5.4	24.5	54.3	79.3	90.0	94.2	96.2
$s_A = 1.00$	$s_B = 0.45$	0.55	0.67	0.82	1.00	1.22	1.49	1.82	2.23
θ_H	0.5	0.8	1.6	4.9	47.9	93.4	96.4	98.7	99.1
$\theta_{3,4,7,8}$	0.4	0.7	1.5	4.8	46.6	93.0	97.4	98.4	98.8
$\theta_{5,6,7,8}$	0.7	1.1	2.0	5.7	48.7	94.0	97.9	98.8	99.2
$s_A = 2.01$	$s_B = 0.20$	0.25	0.30	0.37	0.45	0.55	0.67	0.82	1.00
θ_H	7.4	9.8	14.5	24.8	46.1	72.9	92.9	97.8	99.0
$\theta_{3,4,7,8}$	4.2	6.3	10.8	21.5	45.2	74.4	93.5	97.7	98.7
$\theta_{5,6,7,8}$	5.2	7.6	12.5	23.9	48.0	76.5	94.4	98.1	99.0

[a] $\sigma = 5 \times 10^{-4}$.

$P_{5,6,7,8}(j)$. Incidentally, the residues in proteins analyzed by the X-ray diffraction method are classified by most experimentalists as helical or nonhelical, depending on whether or not their values of ϕ and ψ are in the small range characteristic of the right-handed α-helical conformation. In Figure 2, we can see that (a) the beginning and ending of helical sections are easily located by the probability $P_3(j)$ and $P_7(j)$ (for the beginning), and $P_2(j)$ and $P_6(j)$ (for the ending); it may be noted that, for every peak in P_3 (indicating the start of a helix), there is a peak in P_2 (indicating its termination), and likewise for P_7 and P_6. In Figure 3, we see that (b) $P_H(j)$ and $P_{3,4,7,8}(j)$ are similar to each other, the curve of the latter being somewhat smoother than that of the former, and (c) the curve of $P_{5,6,7,8}(j)$ is shifted slightly toward the C terminus of the polypeptide chain by a few residues compared to that of $P_{3,4,7,8}(j)$. Point c is easily understood from the difference in the definitions of a helical residue, corresponding to the use of $P_{5,6,7,8}(j)$ and $P_{3,4,7,8}(j)$, respectively, as the helix probability.

In Table VII, the average helix content, as defined in three different ways, is calculated and presented for various sets of values of s_A and s_B (for $N = 1000$). They are defined as follows.

$$\theta_H = \frac{1}{N}\sum_{j=1}^{N} P_H(j) \qquad (32)$$

$$\theta_{3,4,7,8} = \frac{1}{N}\sum_{j=1}^{N} P_{3,4,7,8}(j) \qquad (33)$$

$$\theta_{5,6,7,8} = \frac{1}{N}\sum_{j=1}^{N} P_{5,6,7,8}(j) \qquad (34)$$

The second one, $\theta_{3,4,7,8}$, is the fraction of residues whose CO group is involved in the formation of a hydrogen bond, or the ratio of the number of hydrogen bonds formed to the degree of polymerization of the polypeptide. The third one, $\theta_{5,6,7,8}$, is the fraction of residues in the h state. The first one, θ_H, cannot be interpreted directly in terms of the molecular structure of a polypeptide, because it is based on the 2 × 2 matrix formulation or the nearest-neighbor-interaction model, which, when applied to the helix–coil transition in copolymers, does not have a direct relationship with the molecular structure of a polypeptide.

One can see from Table VII that, for the range of s_A and s_B values considered, the differences between $\theta_{3,4,7,8}$, $\theta_{5,6,7,8}$, and θ_H are invariably less than 5% absolute, the usual uncertainty in the determination of α_R-helix content in polypeptides from ORD and CD measurements. Consequently, these results demonstrate that the 2 × 2 matrix formulation is indeed a good approximation even when applied to binary copolymers, if an error of 5% in θ can be tolerated.

Although it has been shown that the 11 × 11 matrix method is not really necessary, at the present time, for the calculation of the mean helix content of a polypeptide copolymer, the application of this method toward defining more precisely the state of a particular amino acid residue in a denatured protein is demonstrated in the next section.

IV. Application to Proteins

In a previous paper,[11] we showed that a good correlation exists between the propensity for a residue to be in the α_R-helical conformation in a denatured protein, as calculated by the 2 × 2 matrix formulation, and the occurrence of the same residue in a helical region in the native structure of that protein. The question then arises as to whether a similar correlation exists between the tendency of a residue to be in one of the eight states (listed in section I) in the denatured protein and the state found for each amino acid residue in the native structure of a particular protein. The purpose of this section is to illustrate the utility of the additional information afforded by the 11 × 11 matrix which enables us to investigate whether such a correlation exists.

The probability that the jth residue of a protein in the denatured condition is in state i, where $i = 1, \ldots, 8$, is given by eq 26. Numerical values[19] for the helix nucleation and helix propagation parameters, σ and s, respectively, for each

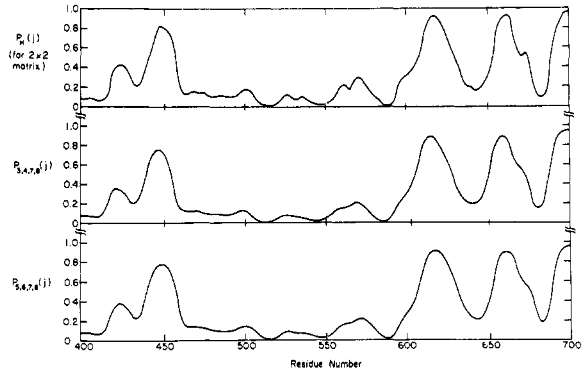

Figure 3. Curves of $P_{3,4,7,8}(j)$, $P_{5,6,7,8}(j)$, and $P_H(j)$ for j = 400–700, *i.e.*, for residues as in Figure 2. The ordinate has the units of normalized probability and the abscissa has units of chain site j. $s_A = 2.01$, $s_B = 0.50$, and $\sigma = 5 \times 10^{-4}$.

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

Table VIII
Assignment of Amino Acid Residues to Three Categories[a] According to Helix-Forming Power[b]

Helix breaker (s = 0.385)	Helix indifferent (s = 1.00)	Helix former (s = 1.05)
Pro	Lys	Val
Ser	Tyr	Gln
Gly	Asp	Ile
Asn	Thr	His
	Arg	Ala
	Cys	Trp
	Phe	Met
		Leu
		Glu

[a] $\sigma = 5 \times 10^{-4}$ for all residues. [b] See ref 19.

of the 20 naturally occurring amino acid residues are given in Table VIII. They correspond to an assignment of each amino acid residue to one of three categories reflecting the helix-forming capacity of the residue in question, *viz.*, (i) helix breaker, (ii) helix indifferent, and (iii) helix former. These assignments and the choices of the parameters are discussed elsewhere.[11,19] The assignments given in Table VIII were used in the calculation of the probabilities discussed in the following two examples.

(a) Sperm Whale Myoglobin. In its native structure, the globular protein sperm whale myoglobin has eight α_R-helical sequences, several of which are separated by only one residue in the coil state; *i.e.*, residues at positions 20, 36, and 58 are in state 4 according to the non-α_R-helical (ϕ, ψ) angles[20] for these residues. The existence of such abrupt breaks in essentially continuous helical regions can be explained ɔnly in terms of long-range interactions.[21] Nevertheless, it might be argued that, if the helix must be disrupted for the sake of the overall conformational stability of the native structure, then it will "break" at the weakest point, *viz.*, at the residue whose probability for being in state 4 in the denatured condition is the greatest in that region of the chain. The probabilities $P_{5,6,7,8}(j)$ and $P_4(j)$ for myoglobin are shown in Figure 4; the former represents the probability that a residue has (ϕ, ψ) angles corresponding to those of an α_R helix, while the latter represents the probability that the CO and NH groups are both hydrogen bonded while the intervening (ϕ, ψ) angles are free to adopt any value. Even though the probability that a residue is in state 4 is very small ($\sim 3 \times 10^{-5}$) in the denatured condition, relative maxima in the various regions are found at residues 22, 38, 48, and 55, in close proximity to the breaks found in the native structure at positions 20, 36, and 58. This information could not possibly have been obtained with the 2 × 2 matrix formulation, as was already noted in section II. The comparatively high probability of occurrence of α_R helix in the denatured state,[11] indicated by $P_{5,6,7,8}(j)$, also corresponds fairly well with the α_R-helical sections found in the native structure (indicated by horizontal bars in Figure 4). It should be noted that no long-range interactions have been included in eq 26; all correlations are the result of near-neighbor interactions along the protein chain. Once experiments on copolymer systems, now in progress in this laboratory,[21] have provided more precise values of σ and s for each amino acid residue, it is expected that even better correlations will be established.

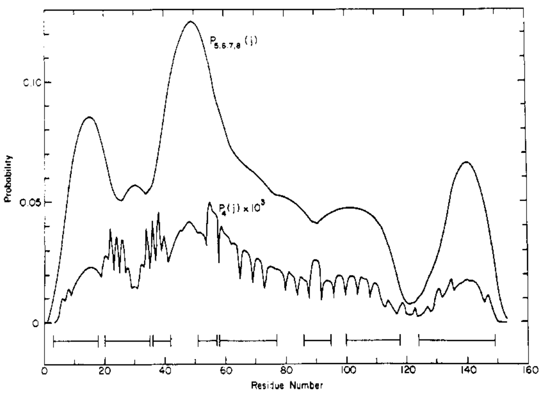

Figure 4. Curves of $P_{5,6,7,8}(j)$ and $P_4(j)$ for denatured sperm whale myoglobin. The ordinate has units of normalized probability and the abscissa has units of chain site. Values of s are given in Table VIII. The horizontal bars at the bottom denote those regions of the native protein found to be in the α_R conformation by X-ray diffraction analysis.

(b) Hen Egg White Lysozyme. In contrast to myoglobin, lysozyme has only 36% helix content in its native structure. The precise position on the lysozyme chain at which an α_R helix is likely to start and end can be determined in several ways. First, one can calculate from the 2 × 2 matrix formulation the quantities $P_{ch}(j)$ and $P_{hc}(j)$, described in section II, which correspond to the probabilities that the jth residue is in state h preceded by a residue in state c (beginning of helical section) and that the jth residue is in state c preceded by a residue in state h (end of helical section), respectively. These probability profiles for the beginning and ending of helical sections are shown in the upper part of Figure 5. Alternatively, the 11 × 11 matrix yields the probabilities $P_2(j)$ and $P_3(j)$, which also represent the likelihood for beginning and ending helical regions along the chain. The corresponding profiles are shown in the central part of Figure 5. From an inspection of these profiles, shown in the middle of Figure 5, it can be seen that for every solid peak (beginning of helical section) there is a corresponding dashed peak (end of helical section). In principle, a helix can exist between each pair of peaks, the higher the peak the more likely the existence of a helix. Thus, in terms of states defined by the h and c notation of the 2 × 2 matrix formulation, positions 5–16, 28–36, 94–100, 107–113, and 118–127 are the most likely ones to be α_R helical in the native structure, on the basis of near-neighbor interactions. In a similar manner for the $P_2(j)$, $P_3(j)$ profiles, the residues at positions 5–15, 28–35, 94–99, 107–112, and 118–129 are the most probable ones to be α_R helical. Those cases in which the two peaks are less than three residues apart are neglected.

The most probable positions for the beginning and ending of helical sections can also be determined from the profiles of $P_7(j)$ and $P_6(j)$, respectively. These are shown in the lower part of Figure 5, and are much more distinct than those for the other two probability sets, *i.e.*, $P_{ch}(j)$, $P_{hc}(i)$ and $P_2(j)$, $P_3(j)$. From the $P_6(j)$ and $P_7(j)$ profiles, residues at positions 7-13, 30-33, 95-98, 108-114, and 117-126 are the most probable ones to be α_R helical. Comparing the three methods illustrated in Figure 5, it can be seen that the only difference is the definition of those residues which one considers to be helical. Since agreement with the positions of the native α_R-helical sections[22] located at 5-15, 25-36, 88-99, 109-114,

[20] H. C. Watson, *Progr. Stereochem.*, **4**, 299 (1969).

[21] H. A. Scheraga, *Chem. Rev.*, **71**, 195 (1971).

[22] C. C. F. Blake, D. F. Koenig, K. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, *Nature (London)*, **206**, 757 (1965); *Proc. Roy. Soc., Ser. B*, **167**, 365 (1967).

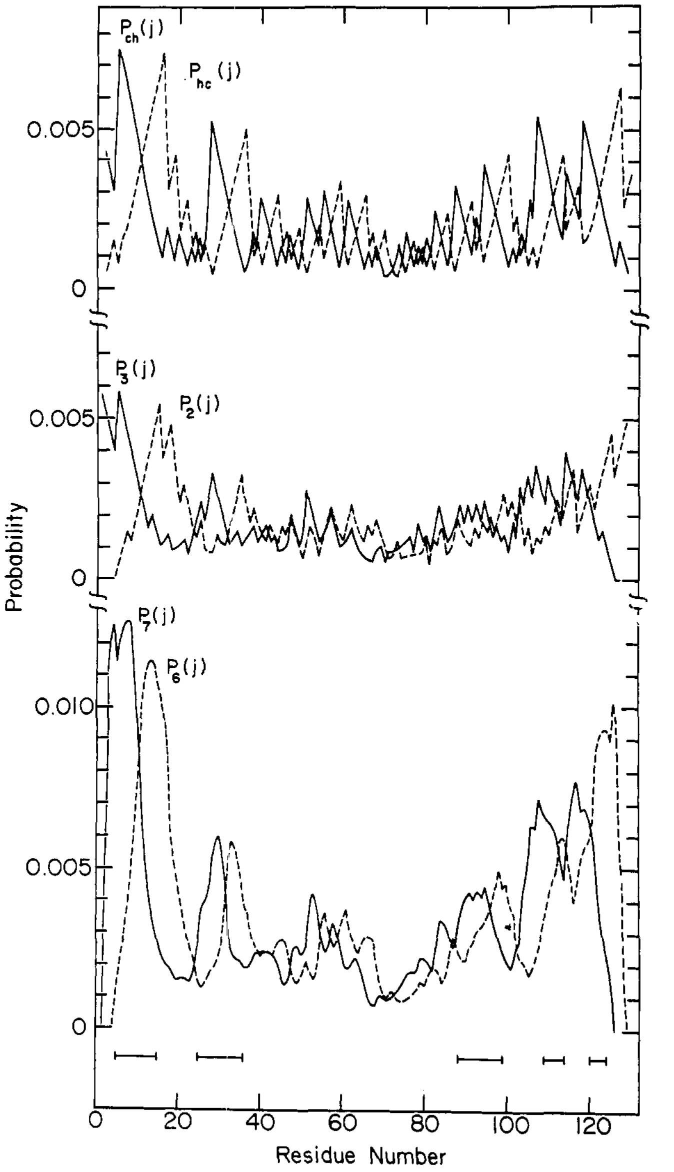

Figure 5. Curves of normalized probabilities for denatured hen egg white lysozyme. The ordinate has units of normalized probability and the abscissa has units of chain site. Values of s are given in Table VIII. The horizontal bars at the bottom denote those regions of the native protein found to be in the α_R conformation by X-ray diffraction analysis.

120–124 is very good, the problem (at least for this protein) is how to define a cutoff probability below which a residue will be predicted to be in a nonhelical conformation in the native structure.

One possible solution[23] to the cutoff problem is to combine the above criterion (*e.g.*, a helix is bound by peaks in P_3 and P_2) with the one used in ref 11 and 19 such that a residue is predicted to be in the α_R-helical conformation in the native structure of a particular protein if first, its helix probability exceeds the mean helix probability for the protein (θ_H from the 2 × 2 matrix formulation and $\theta_{3,4,7,8}$, $\theta_{5,6,7,8}$ from the 11 × 11 matrix formulation) and second, it lies within a certain region bounded by the helix beginning and ending curves such as $P_{ch}(j)$, $P_{hc}(j)$, or $P_3(j)$, $P_2(j)$, or $P_7(j)$, $P_6(j)$ as discussed above. Applying these two criteria to lysozyme, α_R helix is predicted to occur at positions 5–15, 29–34, 94–98, 107–112, and 118–124 by the 2 × 2 matrix model and at positions 5–15, 28–35, 94–99, 107–112, and 118–126 by the 11 × 11 matrix model, both of which agree well with the experimentally determined helix positions given above.

TABLE IX
ALL POSSIBLE TRIPLETS AND QUADRUPLETS IN THE ZIMM–BRAGG MODEL[a]

$i-3$	$i-2$	$i-1$	i
0	0	0	0, 1
1	0	0	0, 1
0	1	0	0
1	1	0	0
0	0	1	0, 1
0	1	1	0, 1
1	1	1	0, 1

[a] See ref 3.

The problem of predicting the conformational states for those residues whose helix probability lies near the mean helix probability (the cutoff question as described above) might also be resolved if it can be shown that those particular residues exhibit a pronounced preference for either adopting or rejecting an alternative conformation, *e.g.*, a β structure. This question is currently under investigation.

In this example, it has been shown that the 2 × 2 and 11 × 11 matrix formulations are nearly equivalent, as far as the location of helical sections is concerned. This arises primarily from the small number of parameters (only three) used to characterize the helix-forming capacity of the 20 naturally occurring amino acids, as described in Table VIII. It is quite likely that, once precise parameters have been determined for all amino acids, significant differences will occur between the two methods, *i.e.*, the 2 × 2 and 11 × 11 matrix formulations, as used above.

V. Conclusion

From the numerical examples of section III, we have shown that the average properties of copolymers, calculated by the 11 × 11 matrix method developed here, differ only slightly from those of the simpler and less physically realistic 2 × 2 matrix formulation. Thus, under the conditions described in section III, one is justified in using the 2 × 2 matrix method for the analysis of experimental data on helix–coil transitions in specific-sequence copolymers. However, the information content of the 11 × 11 matrix is larger than that of the 2 × 2 or even the 8 × 8 matrix method of Zimm and Bragg.[3] Thus, it would be useful to apply this larger matrix to known protein sequences to investigate the possibility of correlating or even predicting specific features of the three-dimensional structure in the native and denatured states, as shown here by two examples.

Appendix A. Contraction of the Zimm–Bragg 8 × 8 Matrix

In this appendix, we present a systematic procedure for reducing the 8 × 8 statistical weight matrix of Zimm and Bragg[3] to a 4 × 4 matrix.[7] All possible triplets and quadruplets are listed in Table IX in terms of 0's and 1's, the two possible states per residue. Since triplet 101 is an impossible one in the model, there are only seven possible triplets, and we can therefore start with a 7 × 7 matrix[24] rather than

[23] P. N. Lewis and H. A. Scheraga, *Arch. Biochem. Biophys.*, **144**, 584 (1971).

[24] Nagai[4] also used a 7 × 7 matrix. However, he defined the states of a residue in a manner similar to that of Lifson and Roig,[6] and he assumed that states chc and chhc did not occur. The difference between the Zimm–Bragg and Lifson–Roig models is considered in Appendix B.

Table X
7 × 7 Matrix V for the Zimm–Bragg Model[a] with $\mu = 3$

					α	β	γ		δ		
				i	0	0	0	0	1	1	1
				$i-1$	0	0	1	1	0	1	1
	$i-3$	$i-2$	$i-1$	$i-2$	0	1	0	1	0	0	1
α	0	0	0		1				σs		
β	1	0	0		1				0		
γ	0	1	0			1					
	1	1	0			1					
δ	0	0	1				1			s	
	0	1	1					1			s
	1	1	1					1			s

[a] See ref 3.

with the 8 × 8 matrix that Zimm and Bragg treated originally. The 7 × 7 matrix V is given in Table X. As in section II, a blank is a nonsense zero, and is distinguished from a zero for a disallowed quadruplet (*i.e.*, 1001) which is shown explicitly in Table X. There are 11 nonvanishing matrix elements, corresponding to the 11 possible quadruplets. The partition function Z for the Zimm–Bragg model with $\mu = 3$ for a copolymer of N amino acids, introduced in section II, is given by

$$Z = \mathbf{s} \prod_{j=5}^{N} \mathbf{V}_{\mathrm{A}(j)} \mathbf{t} \tag{A-1}$$

where $\mathbf{V}_{\mathrm{A}(j)}$ is the 7 × 7 matrix of Table X with the values of the elements corresponding to amino acid type A(j), and **s** and **t** are the following row and column vectors, respectively.

$$\mathbf{s} = (1, 0, 0, 0, 0, 0, 0) \tag{A-2}$$

and

$$\mathbf{t} = (1, 1, 1, 1, 1, 1, 1)^{+} \tag{A-3}$$

where the plus sign designates a transpose. The fact that the first four amino acid residues of the chain are always in state 0 (see footnote 15 in section II) is reflected in the form of the vector **s** and in the initial index of j as 5 in eq A-1. In Table X, the classification of the seven possible triplets into four groups, and the corresponding division of the 7 × 7 matrix V into 4^2 or 16 submatrices, are shown. All submatrices are either of the two types discussed in section II. Therefore, eq 27 can be derived from eq A-1 by replacing each submatrix by a number. Since the rank of the resulting contracted matrix, eq 28, equals its dimension, then no further contraction can be performed even for the case of a homopolymer. This is in direct contrast to the 44 × 44 matrix which can be contracted to an 11 × 11 matrix for a copolymer, and further to a 4 × 4 matrix for a homopolymer.

Appendix B. Application of the Lifson–Roig Model

(I) Comparison of the Zimm–Bragg and Lifson–Roig Models for Homopoly(amino acids). The helix–coil transition of a homopoly(amino acid) is usually analyzed by either the Zimm–Bragg[3] or Lifson–Roig[6] models. The principal difference between these two models is in the description of the state of a residue. The Zimm–Bragg model focuses attention on the hydrogen bondedness of the N–H of the ith residue to the C=O of the ($i - 4$) residue, while the Lifson–Roig model makes use of the condition of the dihedral (ϕ, ψ) angles of each amino acid. In both models, the state of the ith amino acid is correlated with those of its nearby neighbors by means of a correlation matrix, 4 × 4 for the Zimm–Bragg model and 3 × 3 for the Lifson–Roig model. Lifson and Roig[6] have pointed out that their model is more exact (in the sense that it describes the physical situation more accurately) than the model of Zimm and Bragg. The origin of this deficiency in the Zimm–Bragg model as well as the relationships among the parameters used in the two models are given in this appendix.

The ith residue is considered to be in a coil state according to the Zimm–Bragg model, if the N–H of the ith residue is *not* hydrogen bonded to the C=O of the ($i - 4$) residue. This state is assigned a statistical weight of unity. On the other hand, a coil state in the Lifson–Roig model is one in which the dihedral (ϕ, ψ) angles of an amino acid residue are not those characteristic of an α_{R} helix; a statistical weight of u is assigned to this state. In addition, in the Lifson–Roig model, a statistical weight of v is assigned to a residue whose (ϕ, ψ) angles are those characteristic of an α_{R} helix but which is not located in the interior of an α_{R} helix. Thus, for a single residue not involved in an α_{R} helix and whose (ϕ, ψ) angles may take on any values, the Lifson–Roig and Zimm–Bragg statistical weights are ($u + v$) and 1, respectively, which are identical, *viz.*

$$u + v = 1 \tag{B-1}$$

As an example, for the case in which three consecutive residues are not constrained by a helical hydrogen bond, the Zimm–Bragg statistical weight is given by 1^3 which is the same as

$$(u + v)^3 = u^3 + 3u^2v + 3uv^2 + v^3 \tag{B-2}$$

From this expression, it should be noted that the situation in which three consecutive residues have α_{R}-helical (ϕ, ψ) angles but no hydrogen bond (represented by the term v^3 in eq B-2) is also counted. Clearly this is incorrect; the correct statistical weight for this conformation is v^2w, where the statistical weight w reflects the formation of a hydrogen bond. Consequently, the Zimm–Bragg model *in effect* overcounts the states accessible to the poly(amino acid) chain; *i.e.*, in this example, the *statistical weight for all possible conformations* of the poly(amino acid) triplet is really given by expression B-3 rather than by the Zimm–Bragg expression B-4.

$$u^3 + 3u^2v + 3uv^2 + v^2w \tag{B-3}$$

$$u^3 + 3u^2v + 3uv^2 + v^3 + v^2w \tag{B-4}$$

Since the numerical value of v for poly(amino acids) is about 10^{-2}, the contributions of terms such as v^3, v^4, etc., to the partition function for the Zimm–Bragg model are negligible. Thus, as Lifson and Roig[6] have pointed out, if the statistical weight for the coil state in the Zimm–Bragg model is taken as ($u + v$), which is set equal to unity as indicated in eq B-1, then appropriate average quantities calculated from the partition function of either model will be the same to a very good approximation.

The statistical weight [for homopoly(amino acids)] assigned to a helical sequence of j units in the Zimm–Bragg model is σs^{j-2}, where $j \geq 3$, while the corresponding quantity in the Lifson–Roig model is v^2w^{j-2}. These quantities can be equated *only* if the same reference state is chosen for both models. If we choose the coil state of the Zimm–Bragg model as the reference state, then $\sigma s^{j-2}/1^j = v^2w^{j-2}/(u + v)^j$, which leads to the following identities.

$$\sigma^{1/2} = v/(u + v) \qquad s = w/(u + v) \tag{B-5}$$

In practice, the coil state in the Lifson–Roig model, which is characterized by nonhelical (ϕ, ψ) angles, is assigned a

Table XI
Statistical Weights for the Lifson–Roig Model

State	(ϕ, ψ)[a]	Hydrogen bond on[b] CO	Hydrogen bond on[b] NH	Statistical weight[c] Real	Statistical weight[c] Dummy
1	c	No	No	u	p_1
2	c	No	Yes	$v^{-1/2}w^{1/2}u$	p_2
3	c	Yes	No	$v^{-1/2}w^{1/2}u$	p_3
4	c	Yes	Yes	$v^{-1}wu$	p_4
5	h	No	No	v	p_5
6	h	No	Yes	$v^{1/2}w^{1/2}$	p_6
7	h	Yes	No	$v^{1/2}w^{1/2}$	p_7
8	h	Yes	Yes	w	p_8

[a] The symbol h corresponds to the states of the dihedral angles ϕ and ψ of a residue in which they are those characteristic of an α_R helix; if the values of ϕ and ψ are *not* those of an h state, then the residue is in a c state. The difference between the definition of a c state in the Lifson-Roig and Zimm–Bragg models (see footnote *a* of Table I) should be noted. [b] The existence or absence of a hydrogen bond on the CO and NH groups is designated by yes or no, respectively. [c] The method of assignment of these eight statistical weights is entirely analogous to those given in Table I; however, the values of the statistical weights for each state in this table are not the same as those in Table I (see eq B-6 and B-7 for the appropriate conversion expressions).

statistical weight u of unity so that the following relationships result

$$\sigma^{1/2} = v/(1 + v) \quad \text{or} \quad v = 1/(\sigma^{-1/2} - 1) \qquad \text{(B-6)}$$

$$s = w/(1 + v) \quad \text{or} \quad w = s/(1 - \sigma^{1/2}) \qquad \text{(B-7)}$$

where the v and w in eq B-6 and B-7 are really v/u and w/u, since u is taken as the reference state.

The partition functions for the Zimm–Bragg (ZB) and Lifson–Roig (LR) models are given by eq B-8 and B-9, respectively.

$$Z_{\text{ZB}}(\sigma, s, N) = (1, 0, 0, 0)\begin{bmatrix} 1 & 0 & 0 & \sigma s \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & s \end{bmatrix}^{N-4}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad \text{(B-8)}$$

$$Z_{\text{LR}}(u, v, w, N) = (0, 0, 1)\begin{bmatrix} w & v & 0 \\ 0 & 0 & u \\ v & v & u \end{bmatrix}^{N}\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \qquad \text{(B-9)}$$

In order to demonstrate the effect of *improperly* equating $s = w$ and $\sigma^{1/2} = v$, consider a homopolymer of $N = 100$, $s = 1.0$, and $\sigma^{1/2} = 0.01$. The values[25] of θ_{H}, the mean helix content calculated from the appropriate derivatives of Z of eq B-8 and B-9, are 0.104 and 0.077, respectively, for the improper conversion of s and σ. The value of θ_{H} from eq B-9 is 0.114 (which should be compared with 0.104 from eq B-8) when the proper conversions given in eq B-6 and B-7 are used. For the case of $N = 1000$, the three corresponding values of θ_{H} are 0.443, 0.246, and 0.445, respectively. These results demonstrate that, for all practical purposes, the Zimm–Bragg and Lifson–Roig models are equivalent for homopolymers, once the proper conversion between the parameters σ, s and v, w (given by eq B-6 and B-7) is recognized.

(II) Lifson–Roig Model for Specific-Sequence Copolymers. In part I of this appendix, one of the fundamental differences between the Zimm–Bragg and Lifson–Roig models for the helix–coil transition in *homopolymers* was shown to be that, in the Zimm–Bragg model, states such as three consecutive residues with α_R-helical (ϕ, ψ) angles but no hydrogen bond are effectively (but incorrectly) assumed to occur. This overestimate of the number of conformations available to the poly(amino acid) chain is also present in the 11 × 11 correlation matrix (Table V) whose statistical weights are based on the Zimm–Bragg model. However, these incorrect conformations (as discussed in part I of this appendix) contribute factors of the order of v^3 and less; therefore, these effects on the average properties of specific-sequence copolymers are certainly less than the side chain–side chain interactions neglected by both models. In light of this, either the Zimm–Bragg or Lifson–Roig model provides an adequate description of the physical system. However, for the sake of completeness, the Lifson–Roig model for specific-sequence copolymers is pre-

Table XII
All Possible Triplets and Quadruplets (Lifson–Roig Model)

$i-3$	$i-2$	$i-1$	i	$i-3$	$i-2$	$i-1$	i	$i-3$	$i-2$	$i-1$	i
1	1	1	1, 3, 5	1	1	5	1, 3, 5	7	6	6	6
2	1	1	1, 3, 5	2	1	5	1, 3, 5	8	6	6	6
5	1	1	1, 3, 5	5	1	5	1, 3, 5	7	7	5	6
5	2	1	1, 3, 5	5	2	5	1, 3, 5	7	7	6	6
6	2	1	1, 3, 5	6	2	5	1, 3, 5	7	8	6	6
1	5	1	1, 3, 5	1	3	5	5	8	8	6	6
2	5	1	1, 3, 5	2	3	5	5	1	3	7	5, 7
5	5	1	1, 3, 5	5	3	5	5	2	3	7	5, 7
5	5	2	1, 3, 5	5	4	5	5	5	3	7	5, 7
5	6	2	1, 3, 5	6	4	5	5	5	4	7	5, 7
6	6	2	1, 3, 5	1	5	5	1, 3	6	4	7	5, 7
1	1	3	5, 7	2	5	5	1, 3	3	7	7	5, 7
2	1	3	5, 7	3	5	5	5	4	7	7	5, 7
5	1	3	5, 7	4	5	5	5	7	7	7	6, 8
5	2	3	5, 7	5	5	5	2, 4	7	7	8	6, 8
6	2	3	5, 7	5	5	6	2, 4	7	8	8	6, 8
1	5	3	5, 7	5	6	6	2, 4	8	8	8	6, 8
2	5	3	5, 7	6	6	6	2, 4				
5	5	3	5, 7	3	7	5	5				
5	5	4	5, 7	4	7	5	5				
5	6	4	5, 7	7	5	5	6				
6	6	4	5, 7	7	5	6	6				

[25] $\theta_{\text{H}} = (\partial \ln Z)/(\partial \ln \delta)$; $\delta = s$ and w for the Zimm–Bragg and Lifson–Roig models, respectively.

Table XIII
61 × 61 Matrix V

				α											β											γ					δ				
			i	1	1	1	1	1	1	1	1	2	2	2	3	3	3	3	3	3	3	3	4	4	4	5	5	5	5	5	5	5	5	5	5
			i − 1	1	1	1	2	2	5	5	5	5	6	6	1	1	1	2	2	5	5	5	5	6	6	1	1	1	2	2	3	3	3	4	4
			i − 2	1	2	5	5	6	1	2	5	5	5	6	1	2	5	5	6	1	2	5	5	5	6	1	2	5	5	6	1	2	5	5	6
	i − 3	i − 2	i − 1																																
α	1	1	1	p_1											p_3											p_5									
	2	1	1	p_1											p_3											p_5									
	5	1	1	p_1											p_3											p_5									
	5	2	1		p_1											p_3											p_5								
	6	2	1		p_1											p_3											p_5								
	1	5	1			p_1											p_3											p_5							
	2	5	1			p_1											p_3											p_5							
	5	5	1			p_1											p_3											p_5							
	5	5	2				p_1											p_3											p_5						
	5	6	2					p_1											p_3											p_5					
	6	6	2					p_1											p_3											p_5					
β	1	1	3																												p_5				
	2	1	3																												p_5				
	5	1	3																												p_5				
	5	2	3																													p_5			
	6	2	3																													p_5			
	1	5	3																														p_5		
	2	5	3																														p_5		
	5	5	3																														p_5		
	5	5	4																															p_5	
	5	6	4																																p_5
	6	6	4																																p_5
γ	1	1	5						p_1											p_3															
	2	1	5						p_1											p_3															
	5	1	5						p_1											p_3															
	5	2	5							p_1											p_3														
	6	2	5							p_1											p_3														
δ	1	3	5																																
	2	3	5																																
	5	3	5																																
	5	4	5																																
	6	4	5																																
ε	1	5	5								p_1											p_3													
	2	5	5								p_1											p_3													
ζ	3	5	5																																
	4	5	5																																
η	5	5	5									p_2											p_4												
	5	5	6										p_2											p_4											
	5	6	6											p_2											p_4										
	6	6	6											p_2											p_4										
θ	3	7	5																																
	4	7	5																																
ι	7	5	5																																
	7	5	6																																
	7	6	6																																
	8	6	6																																
κ	7	7	5																																
	7	7	6																																
	7	8	6																																
	8	8	6																																
λ	1	3	7																																
	2	3	7																																
	5	3	7																																
	5	4	7																																
	6	4	7																																
μ	3	7	7																																
	4	7	7																																
ν	7	7	7																																
	7	7	8																																
	7	8	8																																
	8	8	8																																

TABLE XIII *Continued*

				ε		ζ		η				θ		ι				κ				λ					μ		ν			
			i	5	5	5	5	5	6	6	6	5	5	5	6	6	6	5	6	6	6	7	7	7	7	7	7	7	7	8	8	8
			i − 1	5	5	5	5	5	5	6	6	7	7	5	5	6	6	7	7	8	8	3	3	3	4	4	7	7	7	7	8	8
			i − 2	1	2	3	4	5	5	5	6	3	4	7	7	7	8	7	7	7	8	1	2	5	5	6	3	4	7	7	7	8
	i − 3	i − 2	i − 1																													
α	1	1	1																													
	2	1	1																													
	5	1	1																													
	5	2	1																													
	6	2	1																													
	1	5	1																													
	2	5	1																													
	5	5	1																													
	5	5	2																													
	5	6	2																													
	6	6	2																													
β	1	1	3																			p_7										
	2	1	3																			p_7										
	5	1	3																			p_7										
	5	2	3																				p_7									
	6	2	3																				p_7									
	1	5	3																					p_7								
	2	5	3																					p_7								
	5	5	3																					p_7								
	5	5	4																						p_7							
	5	6	4																							p_7						
	6	6	4																							p_7						
γ	1	1	5	p_5																												
	2	1	5	p_5																												
	5	1	5	p_5																												
	5	2	5		p_5																											
	6	2	5		p_5																											
δ	1	3	5			p_5																										
	2	3	5			p_5																										
	5	3	5			p_5																										
	5	4	5				p_5																									
	6	4	5				p_5																									
ε	1	5	5																													
	2	5	5																													
ζ	3	5	5					p_5																								
	4	5	5					p_5																								
η	5	5	5																													
	5	5	6																													
	5	6	6																													
	6	6	6																													
θ	3	7	5											p_5																		
	4	7	5											p_5																		
ι	7	5	5						p_6																							
	7	5	6							p_6																						
	7	6	6								p_6																					
	8	6	6								p_6																					
κ	7	7	5												p_6																	
	7	7	6													p_6																
	7	8	6														p_6															
	8	8	6														p_6															
λ	1	3	7									p_5															p_7					
	2	3	7									p_5															p_7					
	5	3	7									p_5															p_7					
	5	4	7										p_5															p_7				
	6	4	7										p_5															p_7				
μ	3	7	7															p_5											p_7			
	4	7	7															p_5											p_7			
ν	7	7	7																p_6											p_8		
	7	7	8																	p_6											p_8	
	7	8	8																		p_6											p_8
	8	8	8																		p_6											p_8

sented in this appendix. Since most of the details concerned with the formulation of a correlation matrix have been given in sections I and II, only the results will be given here.

The states that a residue can assume and their respective statistical weights are given in Table XI. The only difference between Table I and Table XI is the definition of the c state. In this model, the c state is simply those accessible (ϕ, ψ) dihedral angles other than those in a small region around the α_R-helical angles. Consequently, state 1 in the Zimm–Bragg model (Table I) becomes states 1 and 5 in the Lifson–Roig model (Table XI). States 2, 3, and 4 in the Lifson–Roig model for copolymers *cannot* have α_R-helical (ϕ, ψ) angles; if, for example, the *i*th residue in state 3 has α_R-helical (ϕ, ψ) angles, then the C=O group of the $(i - 1)$ residue will form a hydrogen bond with the N—H of the $(i + 3)$ helical residue which changes the state of the *i*th residue to state 7. The Lifson–Roig parameters u, v, and w are used in Table XI instead of the Zimm–Bragg parameters.

From the definition of the eight states that a residue can assume (Table XI), there are 61 possible triplets and 121 quadruplets, which are given in Table XII. The uncontracted correlation matrix for this model, **V**, is 61 × 61, and is shown in Table XIII; for simplicity, we have omitted *both* kinds of zeros from Table XIII. The end vectors **s** and **t** are given by eq B-10 and B-11, respectively.

$$\mathbf{s} = (1, 0, 0, \ldots, 0, 0) \qquad \text{(B-10)}$$

$$\mathbf{t} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, \ldots, 0, 0)^{+} \qquad \text{(B-11)}$$

The 61 × 61 matrix can be contracted to a 13 × 13 matrix **W** (Table XIV) in a completely analogous fashion to that given for the Zimm–Bragg model. The end vectors **u** and **v** are given by eq B-12 and B-13, respectively.

$$\mathbf{u} = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) \qquad \text{(B-12)}$$

$$\mathbf{v} = (1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)^{+} \qquad \text{(B-13)}$$

Table XIV

Contracted 13 × 13 Matrix W

	α	β	γ	δ	ϵ	ζ	η	θ	ι	κ	λ	μ	ν
α	p_1	p_3	p_5	0	0	0	0	0	0	0	0	0	0
β	0	0	0	p_5	0	0	0	0	0	0	p_7	0	0
γ	p_1	p_3	0	0	p_5	0	0	0	0	0	0	0	0
δ	0	0	0	0	0	p_5	0	0	0	0	0	0	0
ϵ	p_1	p_3	0	0	0	0	0	0	0	0	0	0	0
ζ	0	0	0	0	0	0	p_5	0	0	0	0	0	0
η	p_2	p_4	0	0	0	0	0	0	0	0	0	0	0
θ	0	0	0	0	0	0	0	0	p_5	0	0	0	0
ι	0	0	0	0	0	0	p_6	0	0	0	0	0	0
κ	0	0	0	0	0	0	0	0	p_6	0	0	0	0
λ	0	0	0	0	0	0	0	p_5	0	0	0	p_7	0
μ	0	0	0	0	0	0	0	0	0	p_5	0	0	p_7
ν	0	0	0	0	0	0	0	0	0	p_6	0	0	p_8

Table XV

Description of Sequence-Generating Functions

Sequence-generating function	Description
$H = p_5^3/t^3 + p_5^2p_6p_7/t^4 + p_5p_6^2p_7^2/t_5 + p_6^3p_7^3/t^5(t - p_8)$	Helix
$h = p_5/t + p_5^2/t^2$	Isolated single or double residue(s) with α_R-helical (ϕ, ψ) angles but not constrained by a hydrogen bond
$C = p_1/(t - p_1)$	Coil state
$C_L = p_3/t$	Left-hand end of α_R helix
$C_R = p_2/t$	Right-hand end of α_R helix
$C_4 = p_4/t$	State 4 (Table XI)

In order to obtain the secular equation for this model by the method of sequence-generating functions,[14] a correlation matrix **M** of these functions must be formed (eq B-14 and

$$\mathbf{M} = \begin{array}{ll|ccccc} & i+1 & H & h & C & C_L & C_R \cup C_4 \\ i-1 & i & C_L \cup C_4 & C_R \cup C & C_R \cup h & C_R \cup h \cup C & H \\ \hline C_L \cup C_4 & H & 0 & 0 & 0 & 0 & H \\ C_R \cup C & h & 0 & 0 & h & h & 0 \\ C_R \cup h & C & 0 & C & 0 & C & 0 \\ C_R \cup h \cup C & C_L & C_L & 0 & 0 & 0 & 0 \\ H & C_R \cup C_4 & C_4 & C_R & C_R & C_R & 0 \end{array} \qquad \text{(B-14)}$$

Table XV). The secular equation is then given by $|\mathbf{I} - \mathbf{M}| = 0$. In eq B-14, i, $i - 1$, and $i + 1$ represent blocks to which the sequence-generating functions of Table XV correspond. The expansion of the determinant of the secular equation is given in terms of the sequence-generating functions H, h, C, C_L, C_R, and C_4 by eq B-15. Insertion of the algebraic expres-

$$HC_RC_L(C + 1)(h + 1) - HhCC_4 + HC_4 + hC - 1 = 0 \qquad \text{(B-15)}$$

sions for H, h, C, C_R, C_L, and C_4 given in Table XV into eq B-15 gives the secular equation in terms of the dummy statistical weights $p_1, \ldots, p_8$ (eq B-16). When the real statistical

$$\begin{aligned} &t^{10} - (p_1 + p_8)t^9 + p_1(p_8 - p_5)t^8 + p_1(p_5p_8 - p_5^2)t^7 + \\ &p_5^2(p_1p_8 - p_4p_5)t^6 + [(p_1p_4 - p_2p_3)p_5^3 + (p_5p_8 - p_6p_7)p_4p_5]t^5 + \\ &[(p_1p_4 - p_2p_3)p_5^4 + (p_5p_8 - p_6p_7)p_4p_5p_6p_7 + (p_1p_4 - p_2p_3)(p_5p_8 - \\ &p_6p_7)p_5^2]t^4 + [(p_1p_4 - p_2p_3)p_5^5 + (p_5p_8 - p_6p_7)p_4p_6^2p_7^2 + \\ &(p_1p_4 - p_2p_3)(p_5p_8 - p_6p_7)(p_5p_6p_7 - p_5^3)]t^3 + [(p_1p_4 - \\ &p_2p_3)(p_5p_8 - p_6p_7)(p_6^2p_7^2 + p_5^4 + p_5^2p_6p_7)]t^2 + [(p_1p_4 - \\ &p_2p_3)(p_5p_8 - p_6p_7)(p_6p_7 + p_5^2)p_5p_6p_7]t + (p_1p_4 - p_2p_3)(p_5p_8 - \\ &p_6p_7)p_6^2p_7^2p_5^2 = 0 \qquad \text{(B-16)} \end{aligned}$$

weights given in Table XI are inserted into eq B-16, the familiar Lifson–Roig cubic equation results, *viz.*

$$t^3 - (u + w)t^2 + u(w - v)t + vu(w - v) = 0 \qquad \text{(B-17)}$$

The secular equation for the 13 × 13 correlation matrix $\mathbf{W}_{A(j)}$ shown in Table XIV is simply eq B-16 multiplied by a common factor t^3. As in the 11 × 11 case, this matrix has a lower order rank than its size, however; the similarity transformation for matrix contraction has matrix elements which depend upon the statistical weights and thus, for specific-sequence copolymers, the full-size matrix must be used to form the partition function which is given by

$$Z = \mathbf{u}\prod_{i=1}^{N}\mathbf{W}_{A(j)}\mathbf{v} \qquad \text{(B-18)}$$

Average quantities can be calculated from eq B-18 according to eq 26.

The analogous Lifson–Roig model for the helix–coil transition in specific-sequence copolymers has been given above. Numerical results calculated by the Zimm–Bragg and Lifson–Roig copolymer models (not shown here) are essentially identical.[26] It is interesting to note that, for copolymer systems, a larger correlation matrix[27] is required for the Lifson–Roig model, the reverse situation being the case for homopolymers (3 × 3 *vs.* 4 × 4 for Lifson–Roig and Zimm–Bragg, respectively).

[26] The values of σ and s used in sections III and IV were converted to v and w according to the conversion expressions given by eq B-6 and B-7.

[27] The increased matrix size is due to the existence in the Lifson–Roig model of additional triplets such as 151, 155, . . . , which do not occur explicitly in the Zimm–Bragg model.

Poly(ethyl α-chloroacrylates). Nuclear Magnetic Resonance Spectra and Tetrad Analysis

Bengt Wesslén, Robert W. Lenz, and Frank A. Bovey*

Polymer Science and Engineering Program, Chemical Engineering Department,
University of Massachusetts, Amherst, Massachusetts 01002, and
Bell Laboratories, Murray Hill, New Jersey 07974. Received June 25, 1971

ABSTRACT: Poly(ethyl α-chloroacrylates) of different tacticities have been investigated by nmr spectroscopy at 220 MHz. Tetrad assignments and relative intensities of the tetrads are reported. The results obtained indicate a stereoblock structure of the polymers.

In a previous paper the preparation of ethyl α-chloroacrylate polymers through anionic polymerization reactions was reported.[1] The steric structures of the polymers prepared were investigated with nuclear magnetic resonance spectroscopy. The use of spectra obtained at 60 MHz, as well as at 100 MHz, for the backbone methylene groups led to somewhat inaccurate estimates of the steric configurations of the polymers, owing to poor resolution and overlapping peaks in the spectra. Tacticity differences between polymers prepared under different conditions could be observed qualitatively. However, since it was found that the chemical shifts for the protons of the ethyl ester groups were sensitive to the stereochemistry of the adjacent pseudoasymmetric carbon atoms, quantitative triad analyses were performed by using the resonance signals from the methyl protons of the ethoxy groups. The relative triad intensities thus determined also contained errors because of overlapping peaks.

In the present paper nmr spectra at 220 MHz of the poly(ethyl α-chloroacrylates) are reported. With the increased shift differences realized with the 220-MHz instrument, as compared to 60- and 100-MHz instruments, more highly resolved spectra of the backbone methylene groups of the polymers could be obtained. Tetrad assignments in the spectra have been attempted, and quantitative estimates of the relative tetrad intensities on the basis of these assignments have been carried out.

Results and Discussion

Figure 1 shows nmr spectra at 220 MHz in the backbone methylene region of five samples of poly(ethyl α-chloroacrylate) of different tacticities, ranging from moderately isotactic to predominantly syndiotactic ones, according to the triad analysis reported previously.[1] At 220 MHz, the shift differences for the ethoxy groups of different triads are of approximately the same magnitude as the coupling constant within the ethoxy group. Accordingly, due to overlapping, the proton resonance signals from these groups cannot be used to estimate triad intensities, as was possible at 100 MHz. As shown by the figure, the backbone methylene region of the spectra is comparatively well resolved, although broadly overlapping peaks as well as some background signals complicate the interpretation of these spectra. The background signals can arise from the occurrence of structures having other than the regular head-to-tail arrangement of the ethyl α-chloroacrylate repeating units, because it was shown in the previous paper that a substantial amount of chlorine was lost in the polymerization reactions.[1] Furthermore, end effects might not be negligible since the molecular weights of the polymers in some cases were quite low.[1]

The spectra show a regular progression of the intensities of the different peaks with increasing syndiotacticity in the polymers. The most characteristic feature is the increase in intensity of the singlet at τ 7.09, which is assigned to the completely racemic tetrad, designated *rrr*. A doublet centered at τ 7.24 which decreases in intensity with increasing syndiotacticity is regarded as the upfield doublet of an AB spectrum and assigned to the isotactic *mmm* tetrad. The complete tetrad assignments are shown in Figures 2 and 3. The spectra were interpreted in terms of four AB spectra with geminal coupling constants of 15 Hz, arising from the heterosteric protons of *mmm*, *mmr*, *rmr*, and *mrr* tetrads, and two singlets from the homosteric protons of *rrr* and *mrm* tetrads.[2]

The chemical shifts and coupling constants, as well as the relative intensities of the tetrads, were determined by a curve-resolving technique which utilized computer simulation

[1] B. Wesslén and R. W. Lenz, *Macromolecules*, **4**, 20 (1971).

[2] The notation describing the polymer configurations is that proposed by H. L. Frisch, C. L. Mallows, and F. A. Bovey, *J. Chem. Phys.*, **45**, 1565 (1966). This paper also describes the quantitative relationships between *n*-ads which must be upheld, regardless of stereosequence statistics.