

USE OF RANDOM COPOLYMERS TO DETERMINE THE HELIX-COIL STABILITY CONSTANTS
OF THE NATURALLY OCCURRING AMINO ACIDS

Harold A. Scheraga

Department of Chemistry, Cornell University, Ithaca, New York 14853, USA

Abstract - Short-range interactions dominate in determining the conformational preferences of the amino acid residues in proteins, and a variety of procedures (based on this concept) have been developed to predict the conformational states of the residues of a protein molecule. This paper is concerned with two such conformational states, helical and non-helical (or coil) states; the relative preferences of each of the twenty naturally occurring amino acids for these two states can be expressed in terms of the Zimm-Bragg parameters s and σ . In principle, these parameters can be determined from experimental studies of the thermally-induced helix-coil transition in homopolymers of amino acids in water. However, since most homopolyamino acids are insoluble in water, or are not helical, or (if helical) do not melt between 0 and 100°C in water, resort is had to the host-guest technique in which a random copolymer of a water-soluble host residue and a small amount of a guest residue is prepared. Using the helix-coil transition curves of such a copolymer, and the values of s and σ for the host homopolymer, it is possible to compute the values of s and σ over the temperature range of 0-70°C for the guest residue which is, in turn, each of the twenty naturally occurring amino acids.

In most cases, the random copolymers are prepared from their N-carboxyanhydrides, using suitable blocking groups to protect otherwise-reactive side-chain functional groups. The water-soluble copolymers are checked for the absence of racemization, $\alpha \rightarrow \beta$ shifts, etc., and for the required degree of randomness. For example, for methionine as a guest residue, cyanogen bromide cleavage of the polymer chain yields a series of oligopeptides which indicates that the methionine was incorporated randomly in the chain.

When the resulting values of s and σ are compared to the frequencies of occurrence of helical and non-helical conformations in proteins, a good correlation is obtained in most cases. In those cases where the correlation breaks down, the discrepancy provides information about the influence of specific long-range (e.g., electrostatic) interactions on the helical preferences of the given amino acid residues.

The statistical weights deduced from the one-dimensional short-range interaction model are then incorporated with medium- and long-range interactions into a model to try to predict the three-dimensional structures of globular proteins.

I. INTRODUCTION

In an earlier paper (Ref. 1), it was pointed out that, in order to circumvent the multiple-minimum problem when computing the conformation of a globular protein from its amino acid sequence, it is necessary to adopt approximate procedures to reach the potential energy well in which the native protein lies; then, the approximations can be abandoned, and more exact procedures can be used to reach the minimum of the potential energy well. Such approximate procedures would enable one to avoid the time-consuming and unproductive examination of much of the multi-dimensional energy surface by directing the search to regions of this hypersurface in which the potential energy well of the native protein is likely to lie.

A useful approximate procedure is based on the concept, discussed earlier (Ref. 1), that short-range interactions play a dominant role in determining the conformation of each amino acid residue in a globular protein. This concept underlies currently used short-range interaction models (Ref. 2-6) for "predicting" the conformational states of the residues of a globular protein. The inadequacies of the short-range interaction models are then compen-

sated for by combining them with an algorithm [e.g., a Monte Carlo procedure (Ref. 7)] which introduces medium- and long-range interactions. Such a combined (short-, medium-, and long-range interaction) model (Ref. 7) has the potential for increased accuracy in predicting not only the conformational region but also the actual values of the backbone dihedral angles (Ref. 8) ϕ and ψ of each residue in a globular protein.

In the approximate short-range interaction models, the parameters [e.g., the statistical weights of the matrices in the one-dimensional Ising model treatments (Ref. 6)] can be obtained in either of three ways: (i) from experiments on model polymer systems, (ii) from X-ray data on proteins of known three-dimensional structure, or (iii) by optimization of the data obtained in procedures (i) and/or (ii), i.e., by adjusting the statistical weights to improve the fit (of predicted probabilities of occurrence of various conformational states) to experimental data on globular proteins. In particular, in the Ising model treatments (Ref. 6), statistical mechanical averaging over the whole molecule is carried out in computing the probability that any consecutive number of residues will have a particular conformation, e.g., the right-handed α -helical one. Methods (i)-(iii) consider only short-range interactions, and neglect the longer-range ones.

This paper is devoted to one aspect of the whole problem, viz., to a discussion of method (i), the use of model polymer systems, and will be confined to a consideration of only a portion of the conformational problem, viz., the determination of the Zimm-Bragg (9) parameters s and σ that characterize the relative helix-forming tendencies of the naturally occurring amino acids. Insofar as the results obtained by methods (i) and (ii) agree with each other, we obtain further evidence of the validity of the concept on which this work is based, viz., the dominance of short-range interactions. For those residues for which agreement is not obtained, insight is gained as to the influence of long-range interactions on conformation and as to how and why the short-range interaction model breaks down.

A more detailed discussion of the conformational problem in globular proteins, including a consideration of states other than the helix and (redefined) coil, and definitions of short-, medium-, and long-range interactions, is presented elsewhere (Ref. 10).

II. TWO-STATE MODEL (ALLOWING FOR TWO STATES PER RESIDUE)

In using homopolymers of amino acids to obtain a measure of the helix-forming tendency of each polymer (and hence of each type of amino acid residue) in water, e.g., by studying the thermally-induced helix-coil transition, the data are interpreted in terms of a two state model (the right-handed α -helical conformation, h , and the ensemble of all other conformations, usually designated as the "random coil," c). If experimental "melting" data are available (e.g., data obtained by some optical method that discriminates between helix and coil), these can be fit by adjusting the parameters of a theoretically-computed helix-coil transition curve based, say, on the Zimm-Bragg model (Ref. 9); thus, the parameters are determined. In the Zimm-Bragg model there are three parameters, the chain length (N), which is determined in a separate experiment, and two parameters that are adjusted in fitting the melting data, viz., growth and nucleation parameters (s and σ , respectively). The quantities s and σ thus characterize quantitatively (in the same way that frequencies of occurrence of the helical conformation in globular proteins do) the tendency of each amino acid residue to adopt the helical conformation.

III. USE OF RANDOM COPOLYMERS

For the foregoing approach (using homopolymers) to be feasible experimentally, the homopoly-amino acid must be (i) water soluble, (ii) α -helical, and (iii) capable of being melted in the temperature range between 0 and 100°C. Unfortunately, none of the homopolymers of the naturally occurring amino acids satisfies all three of these requirements. While the use of block copolymers was thought to solve the solubility problem for any amino acid, it has been shown recently (Ref. 11 & 12) that there are serious previously-unrealized problems in the use of this technique; also, the block-copolymer method satisfies requirements (ii) and (iii) only for a few amino acids. Therefore, a technique has been developed which is applicable to *any* amino acid. It involves the study of random copolymers in which the desired amino acid, the "guest," is incorporated at random into a nonionic homopolyamino acid, the "host," which does meet all three requirements (host-guest technique). The helix-coil stability constants (s and σ) of the guest residue can then be determined from its influence on the melting behavior of the host homopolyamino acid, and a knowledge of s and σ of the host homopolymer.

Random copolymers, rather than regular-repeating sequential copolymers of the host and guest residues, have been used because it is easier to synthesize random copolymers and to obtain larger chainlengths. Appropriate theories for analysis of melting data on random copolymers are available (Ref. 13-17). The theoretical and experimental criteria to assess the required degree of randomness in the copolymers are discussed in sections IV and VI.

IV. THEORETICAL MELTING BEHAVIOR OF A RANDOM COPOLYMER

The manner in which the melting behavior of a copolymer of two types of residues differs from that of a homopolymer depends markedly on the relative and absolute values of s and σ of the two types of residues (Ref. 16 & 17). Unlike nucleic acid polymers, amino acid polymers generally exhibit small temperature variations of s , i.e., small values of ΔH , the enthalpy change for the conversion of a coil to a helical state. As a consequence (Ref. 17), the melting curves tend to be relatively insensitive to sequence and, hence, to "small" departures from randomness. This is because local sequence is not important in determining melting behavior. Rather, it is variations in composition on a scale of \bar{L}_h that determine the melting behavior, where \bar{L}_h is the average length of a helical sequence at the transition temperature for an infinite chain. For a homopolymer, \bar{L}_h is $1 + \sigma^{-1/2}$; for a copolymer with similar values of σ , \bar{L}_h has a similar value. Thus, since σ is of the order of 10^{-4} , one must consider variations in composition on the order of 100 residues; i.e., melting tends to occur over regions that are \bar{L}_h units long rather than being localized among individual residues.

The accumulation of a given type of residue into blocks that are 10-20% of \bar{L}_h will not affect the melting behavior (Ref. 15-17), as illustrated in Fig. 1 by the calculated melting curves for regular-sequence copolymers of varying block size. The practical advantage of this is that departures from randomness of this degree, in the synthesis of the "random" copolymer, can be tolerated, since they will not influence the melting behavior of the copolymer.

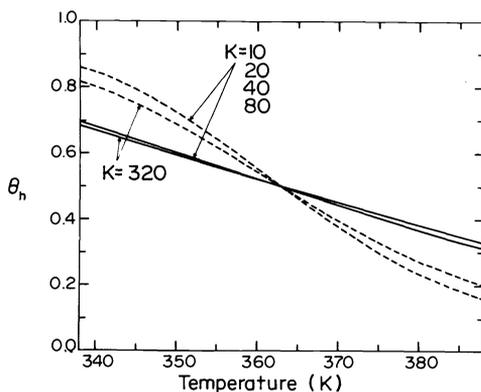


Fig. 1. Theoretical melting curves for infinite chains of regular-sequence copolymers, $\dots[A_{\kappa/2}B_{\kappa/2}]_n\dots$, of varying κ , where the fraction of A residues is 0.5, $\Delta H_B = -200$ cal/mole, $\sigma = 9 \times 10^{-5}$ (----) or 5×10^{-4} (—), $T_A = 342.5$ K, $T_B = 383.5$ K, and $\Delta H_A = \Delta H_B T_A / T_B$ (T_A and T_B are the transition temperatures for the infinite-chain homopolymers). From Ref. 17.

The theory of Lehman and McTague (15) provides an exact treatment of the melting behavior of a copolymer. However, in practice, it is very expensive (in computer time) to use it for analyzing experimental data. Therefore, resort is had to an approximate treatment (Ref. 16), which can be carried out to any desired degree of approximation, p , and the approximate results converge to the exact ones as $p \rightarrow \infty$. The approximate calculations can be performed with much less computer time. In practice, for random copolymers of amino acids, the first one or two approximations ($p = 1$ or 2 , respectively) suffice to give results that agree with those from the theory of Lehman and McTague. The $p = 1$ and 2 approximations correspond to treatments originally presented by Lifson (13) and Allegra (14), respectively.

The theory for the melting behavior of a random copolymer has been tested (Ref. 18) by determining s and σ for two different water-soluble homopolyamino acids and then synthesizing random copolymers of different compositions of these two amino acids. Using the parameters for one of the homopolymers, it was possible to compute those for the other homopolymer by applying the host-guest technique (and associated theory) to the copolymer data. Good agreement was obtained between the parameters computed directly from homopolymer data and those obtained by the host-guest technique, thus establishing the validity of the latter method for use in obtaining the helix-coil stability constants for other amino acids.

V. SYNTHESIS AND CHARACTERIZATION OF COPOLYMERS

Two nonionic homopolymers have served as hosts in these investigations. They are poly[N⁵-(3-hydroxypropyl)-L-glutamine] (PHPG) and poly[N⁵-(4-hydroxybutyl)-L-glutamine] (PHBG). These

are the polymers (Ref. 18) referred to in the last paragraph of section IV. They were prepared (Ref. 18) by polymerizing the N-carboxy anhydride (NCA) of γ -benzyl-L-glutamate to obtain poly(γ -benzyl-L-glutamate) (PBLG) and then converting this material to the final products by substituting the hydroxypropyl or hydroxybutyl group for the benzyl group (using 3-amino-1-propanol or 4-amino-1-butanol, respectively).

Either of these polymers can serve as the host for guest amino acid residues. The choice is dictated by the anticipated values of s and σ for the guest residue. Since PHBG is higher melting than PHPG (in water), the former is used for "helix-breakers" (which shift the transition curve to lower temperatures) and the latter for "helix-makers" (which shift the transition curve to higher temperatures). Thus, the transition curves of the copolymers always appear in a convenient temperature range (10 to 70°C). The copolymers are synthesized from mixtures of the NCA's of the guest residue and γ -benzyl-L-glutamate. The benzyl group is then replaced by reaction with the hydroxyalkylamine to make the polymer water soluble. The guest residues treated thus far are Gly (Ref. 19), Ala (Ref. 20), Ser (Ref. 21 & 22), Leu (Ref. 23), Phe (Ref. 24), Val (Ref. 25), Glu (Ref. 26), Tyr (Ref. 27), Lys (Ref. 28), Asn (Ref. 29), Met (Ref. 30), Arg (Ref. 31), and Asp (Ref. 32). Work on the remaining naturally occurring amino acids is in progress.

For the guest residues which contain a reactive side-chain functional group, the latter has to be protected during the copolymerization of the NCA's of the guest residue and of γ -benzyl-L-glutamate. No protection was needed either during polymerization of the NCA's or during the exchange of the hydroxyalkylamine for the benzyl groups when Gly (Ref. 19), Ala (Ref. 20), Leu (Ref. 23), Phe (Ref. 24), Val (Ref. 25), and Met (Ref. 30) were the guest residues. For the remaining copolymers, protection and deprotection was carried out as follows:

1. *Ser* (Ref. 21)

The serine hydroxyl group was protected with the O-trimethylsilyl group during copolymerization of the NCA's. The benzyl groups of the resulting copolymers were then exchanged with 4-amino-1-butanol. Upon acidification of this product with aqueous acetic acid, the $-\text{Si}(\text{CH}_3)_3$ blocking group was removed.

2. *Glu* (Ref. 26)

The γ -carboxyl groups of the guest and host Glu residues were protected with *tert*-butyl and benzyl groups, respectively, during copolymerization of the NCA's. The *tert*-butyl groups of the resulting copolymers were removed selectively with trifluoroacetic acid, and then the benzyl group was replaced by reaction with 4-amino-1-butanol.

3. *Tyr* (Ref. 27)

The tyrosyl hydroxyl group was protected with the O-tetrahydropyranyl group during copolymerization of the NCA's. The benzyl groups of the resulting copolymers were then exchanged with 3-amino-1-propanol. Upon acidification with aqueous hydrochloric acid, the tetrahydropyranyl protecting group was removed.

4. *Lys* (Ref. 28)

The ϵ -amino group of Lys was protected with the *tert*-butyloxycarbonyl group during copolymerization of the NCA's. The benzyl groups of the resulting copolymers were then exchanged with 4-amino-1-butanol. Upon acidification with aqueous hydrochloric acid, the *tert*-butyloxycarbonyl protecting group was removed.

5. *Asn* (Ref. 29)

The β -amide group of Asn was protected with the (4,4'-dimethoxy)-diphenylmethyl (Mbh) group during copolymerization of the NCA's. The Mbh protecting groups of the resulting copolymers were removed by treatment with trifluoroacetic acid/anisole, and then the benzyl groups were exchanged with 4-amino-1-butanol.

6. *Arg* (Ref. 31)

Copolymers with Arg as the guest residue were obtained from polymers prepared with Orn. The δ -amino group of Orn was protected with the *tert*-butyloxycarbonyl group during copolymerization of the NCA's. The benzyl groups of the resulting copolymers were then exchanged with 4-amino-1-butanol. Upon acidification with aqueous hydrochloric acid, the *tert*-butyloxycarbonyl protecting group was removed. The Orn was then converted to Arg by treatment with O-methylisourea.

7. *Asp* (Ref. 32)

The β -carboxyl group of Asp was protected with the *tert*-butyl group during copolymerization

of the NCA's. The *tert*-butyl groups of the resulting copolymers were removed selectively with trifluoroacetic acid, and then the benzyl group was replaced by reaction with 4-amino-1-butanol.

During the syntheses of all copolymers, care had to be taken to avoid racemization or (in the case of Asp and Asn) $\alpha \rightarrow \beta$ shifts. In the characterization of the resulting copolymers, analyses were carried out to check for complete removal of benzyl and other protecting groups, absence of racemization or $\alpha \rightarrow \beta$ shifts, and absence of deamidation (in the case of Asn). The only copolymers for which significant racemization was detected were those containing L-Asp as the guest residue (Ref. 32); approximately 15% of the Asp residues were found to be in the D-form. Therefore, a correction was applied for the influence of D-Asp to obtain the values of s and σ for L-asp, the correction having been determined from a study of similar host-guest copolymers (containing D-Asp as the guest residue) prepared by the same method.

Fractionation of the resulting water soluble copolymers yielded relatively homogeneous materials, as indicated by values of $\overline{M}_z/\overline{M}_w$ close to unity.

By controlling the relative amounts of the two NCA's, the compositions of the copolymers were arranged to lie between approximately 5% and 15% of the guest residue. Such copolymers were water-soluble, unaggregated, and suitable for study of their thermally-induced helix-coil transitions by optical rotatory dispersion methods.

VI. EXPERIMENTAL DETERMINATION OF "RANDOMNESS"

The theoretical criterion for the required degree of randomness was discussed in section IV. Experimental tests of the degree of randomness of the copolymers were also carried out. In lieu of an actual determination of the nearest-neighbor frequencies or of the kinetics of copolymerization, the copolymers were considered to meet the requirements for randomness demanded by the theory used to compute s and σ if the compositions of the various fractions were independent of chain length (i.e., if the same average percent of guest residue was found in both the short and the long chains). However, in one case (viz., for the copolymers containing methionine as the guest residue), a more rigorous proof of the randomness of the copolymers was provided by analyzing the distribution of fragments from a cyanogen bromide digest of the copolymer (Ref. 33).

For all of the copolymers treated thus far (Ref. 19-32), the requirements for randomness have been met. However, this was not true when Pro was the guest residue. In this case, the Pro NCA polymerizes so rapidly that one obtains essentially poly(L-proline) rather than a random copolymer. Attempts are, therefore, currently being made to obtain the desired values of s and σ for Pro by preparing copolymers by random coupling of blocks of γ -benzyl-L-glutamate and similar blocks containing Pro as an interior residue.

VII. EXPERIMENTAL RESULTS

Using the theory referred to in section IV, the thermally-induced helix-coil transition curves of all of the copolymers were analyzed to obtain s and σ . The experimental data were not accurate enough to obtain precise values of σ or to detect its dependence on temperature. It was, therefore, assumed to be independent of temperature. The corresponding temperature-dependent values of s are shown in Fig. 2. It can be seen that there is a large range of helix-making ($s > 1$) and helix-breaking ($s < 1$) tendencies among the various amino acids. In the case of valine, the residue is helix-breaking at low temperature and helix-making at high temperature, a result which is accounted for by calculations of the helix-coil transition curves from a molecular theory based on empirical potential energy functions (Ref. 34) (see Fig. 3).

On the basis of the data of Fig. 2, we can arrange the various residues on a quantitative scale of s -values at any particular temperature. Such a scale is shown in the second column of Table 1.

It is important to note that the data pertain to neutral pH and, thus, to Glu, Asp, Lys, and Arg in their charged state and to Tyr in its uncharged state. If the state of ionization were altered, the values of s would change significantly. For example, s_{25} for uncharged Glu (Ref. 26) is ≈ 1.3 , compared to the value shown in the second column of Table 1. This difference does *not* arise from electrostatic interactions between charged Glu residues [which, of course, are present in poly(L-glutamic acid) and in copolymers containing a high Glu content]. By keeping the Glu content of the copolymers low (and using salt to suppress inter-residue interactions), it was possible (Ref. 26) to obtain values of s that reflect the *intrinsic* tendency of a Glu residue to be helical (without any influence of other charged residues in the molecule). Within the context of a short-range interaction model, this

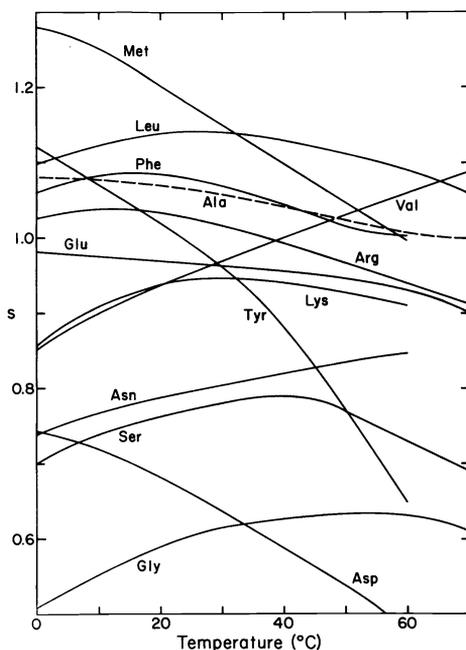


Fig. 2. Experimental s vs. T curves, obtained by the host-guest technique.

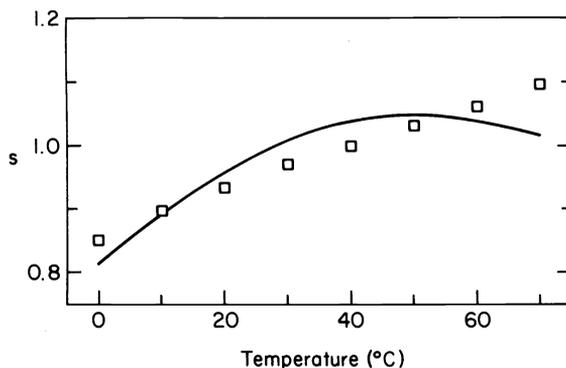


Fig. 3. Comparison of theoretical s vs. T curve for poly(L-valine) in water with experimental results [□]. From Ref. 34. The theoretical curve exhibits the *increase* in s with T shown by the experimental data.

intrinsic tendency reflects interactions between a side chain and its own backbone, and differs for charged and uncharged Glu side chains. Similar differences were obtained (Ref. 32) between charged and uncharged Asp residues. The experiments on Lys (Ref. 28) and Arg (Ref. 31) could not be performed at high enough pH to obtain the behavior of the uncharged residues.

Thus far, all of these experiments have been carried out with either PHBG or PHPG as hosts. In order to determine whether interactions between neighboring residues influence the values of s and σ (i.e., whether medium-range interactions influence the helix-coil stability constants), it would be useful to study the helix-coil transition in other binary water-soluble copolymers, e.g., copolymers of Lys and Ala. One could then determine whether the values of s of Fig. 2 (and the corresponding values of σ) would describe the melting behavior of these copolymers. Such experiments are contemplated.

TABLE 1. Relative helix-forming tendency of several amino acid residues at 25°C at neutral pH

Amino acid	s_{25}			
	From host-guest copolymers	From X-ray data on proteins		
		Uncorrected ^a ($\sigma = 1$)	Corrected ^b ($\sigma = 0.1$)	Corrected ^b ($\sigma = 0.01$)
Met	1.17	1.27	1.08	1.03 ^c
Leu	1.14	1.10	1.03 ^c	1.01 ^c
Phe	1.07	1.00	1.00	1.00
Ala	1.06	1.46 ^c	1.13	1.04
Arg	1.02	1.02	1.01	1.00
Tyr	0.99	0.40 ^c	0.74 ^c	0.91
Glu	0.97	1.61 ^c	1.17 ^c	1.05
Val	0.95	0.70 ^c	0.89	0.96
Lys	0.94	0.96	0.99	1.00
Asn	0.79	0.38 ^c	0.72	0.91 ^c
Ser	0.77	0.44 ^c	0.77	0.92 ^c
Asp	0.66	0.74	0.91 ^c	0.97 ^c
Gly	0.60	0.30 ^c	0.67	0.88 ^c
$\overline{ \Delta }$ ^d		0.25	0.10	0.11
$\overline{\Delta^2}$ ^e		0.11	0.016	0.021

a) From preliminary unpublished data of Isogai et al. (35).

b) Computed by F. R. Maxfield.

c) $|s_{\text{host-guest}} - s_{\text{X-ray}}| > 0.1$ for these residues.

d) $\overline{|\Delta|} = (1/13) \sum (|s_{\text{host-guest}} - s_{\text{X-ray}}|)$.

e) $\overline{\Delta^2} = (1/13) \sum (s_{\text{host-guest}} - s_{\text{X-ray}})^2$.

VIII. COMPARISON WITH X-RAY DATA ON PROTEINS

As pointed out in section I, a quantitative measure of the tendency toward helix formation can also be obtained from X-ray data on proteins of known three-dimensional structure. Relative frequencies of occurrence of interior helical states compared to all other states in proteins are interpreted in terms of the statistical weight s , and a particular set of such data (Ref. 35) (based on 23 proteins) is shown in column 3 of Table 1 [the data in this column were normalized by multiplying all values by a constant factor (selected to make the values for Arg agree), for comparison with the data in the previous column].

Before comparing the data in columns 2 and 3, it should be recognized that computations of values of s from frequencies of occurrence, in this manner, imply the assumption that $\sigma = 1$; i.e., no account is taken of the influence of neighboring residues when computing the frequency of occurrence of any type of residue in an interior helical state (see footnote 45 of Ref. 6a). It is thus inconsistent to use the data of column 3 in a matrix treatment of an Ising model that allows for σ to take on values other than 1. Figure 4, in which the fraction of helical states, θ_h , is plotted against s (using a one-dimensional nearest-neighbor Ising model treatment) for various values of σ , illustrates the problem; it can be seen that a given value of θ_h corresponds to *different* values of s , depending on the value of σ . The curve for $\sigma = 1$ was obtained by computing θ_h [as $s/(1+s)$], using the values of s in column 3 of Table 1. [It should be noted that these computed values of θ_h are *not* the ones observed in proteins because, as stated above, the values of s had been normalized to make the values for Arg in columns 2 and 3 agree.] The curves for $\sigma = 0.1$ and 0.01 were obtained with the Ising model by computing the values of s that would give the same values of θ_h that had been

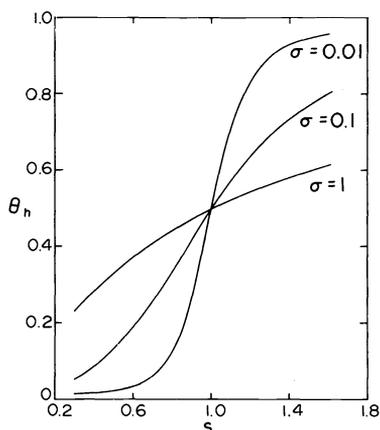


Fig. 4. Dependence of θ_h on s for various values of σ --computed as described in the text.

obtained for $\sigma = 1$. These values are also shown in columns 4 and 5, respectively, in Table 1.

The choice of $\sigma = 0.1$ and 0.01 for these trial calculations can be justified as follows. For the 23 proteins used to obtain the data (Ref. 35) in column 3, the average length of the helical sequences is 8.6 (Ref. 35). This would be the value of \bar{L}_h in a homopolymer if σ were 0.02. Since θ_h in the protein data set is not 0.5 (i.e., the value at the midpoint of the transition in an infinite homopolymer) but closer to 0.3 (Ref. 35), and the protein chains are not infinite, this estimate of $\sigma = 0.02$ is very rough. Thus, the selected values of 0.1 and 0.01 are in a reasonable range.

On the basis of the deviations defined in footnotes (d) and (e) of Table 1, it can be seen that the data in column 4 (for $\sigma = 0.1$) give the best agreement with those in column 2. For this choice of σ , only 4 residues show a deviation $|s_{\text{host-guest}} - s_{\text{X-ray}}| > 0.1$, compared to 6 residues for a choice of $\sigma = 0.01$ [see footnote (c)]. It is of interest that Suzuki and Robson (36), using a much more elaborate procedure to analyze X-ray data from proteins, also obtained a value of $\sigma = 0.1$. Likewise, preliminary results (Ref. 37), using the optimization procedure mentioned in section I as procedure (iii), give roughly this value of σ .

While the agreement, overall, is quite good, the discrepancies that remain provide important insights into the nature of the different types of interactions in the copolymers on the one hand and in proteins on the other. The essential point here is that the host-guest technique provides information about the intrinsic helix-forming tendency of a residue (insofar as it is reflected in s and σ), but the long-range interactions in globular proteins have an influence on these parameters which must be taken into account (e.g., by optimization) before applying the results to globular proteins. Comparison between columns 2 and 4 of Table 1 provides information about long-range interactions in proteins.

An example of the role of long-range interactions is provided by a consideration of Glu. This residue is helix-indifferent in the host-guest copolymers, but one of the strongest helix-makers in globular proteins. An analysis (Ref. 38) of the neighbors of Glu residues in proteins showed that positively charged side chains four residues away from the Glu residue in helical segments in proteins greatly enhance its probability to be helical. Such specific interactions, of course, are not present in the random copolymers. Similar long-range interactions were detected (Ref. 38) for Asp, Lys, and Arg in proteins, but the effects of these interactions on the values of s are smaller for these residues. In the case of the other residues for which discrepancies exist, it appears that specific long-range interactions, present in proteins but not in the random copolymers, combine with the short-range interactions to influence the conformational states of these residues in proteins.

The inclusion of more than two states in a more general model (Ref. 6 & 10), and an optimization of the parameters, would presumably eliminate the necessity for normalization (as was done to make the values of s_{25} for Arg in columns 2 and 3 of Table 1 agree).

IX. USE OF DATA FROM SHORT-RANGE INTERACTION MODELS

The same short-range interaction model that was used for the host-guest copolymers can be applied to native globular proteins--again assuming that short-range interactions dominate.

The probability, $P(i|n|\{\rho\})$, that n residues (beginning with the i^{th}) will be in a conformational state $\{\rho\}$, where $\{\rho\}$ is a sequence of conformational states η_i of the n residues, is (Ref. 6b)

$$P(i|n|\{\rho\}) = Z^{-1} e_{\sim 1} \left[\prod_{j=1}^i W_{\sim j} \right] \left[\prod_{k=i+1}^{i+n-1} \frac{\partial W_{\sim k}}{\partial \ln(m_{\sim k}; \eta_{\sim k-1} \eta_{\sim k})} \right]_{\{\rho\}} \left[\prod_{\ell=i+n}^N W_{\sim \ell} \right] e_{\sim N}^* \quad (1)$$

where $W_{\sim i}$ is a statistical weight matrix (for the i^{th} residue) whose elements are functions of the statistical weights, Z is the partition function

$$Z = e_{\sim 1} \left[\prod_{i=1}^N W_{\sim i} \right] e_{\sim N}^* \quad (2)$$

where $e_{\sim 1}$ and $e_{\sim 2}$ are end vectors, and $m_{\sim k}; \eta_{\sim k-1} \eta_{\sim k}$ is the statistical weight of the conformational state η_k (see Refs. 6 and 10 for further discussion of this model). For the case under consideration here, $\{\rho\}$ is a sequence of helical states, and the elements of $W_{\sim i}$ are functions of s_i and σ_i . The values of $P(i|n|\{\rho\})$ can be used to predict whether any sequence of n residues will be helical or not. The extension of this model to include other states is discussed elsewhere (Ref. 6 & 10).

The probability $P(i|n|\{\rho\})$ can be computed from values of s and σ determined both from the host-guest technique and from X-ray data on proteins. Comparison of the results provides information about medium- and long-range interactions not included in the short-range interaction model. The parameters based on X-ray data (Ref. 6) have been improved recently (Ref. 35) by inclusion of more X-ray data and by a re-definition of interior and end helix states. Also, these parameters are being optimized by procedure (iii) of section I. Computations are now in progress (Ref. 35) to determine whether the predicted locations of helical segments will agree better with experimental data than earlier computations (Ref. 6) based on the preliminary set of statistical weights.

A similar use of the values of s and σ determined by the host-guest technique has been made by Ptitsyn and coworkers (39-41) in their statistical mechanical treatment of globular proteins.

As indicated in section I, the shortcomings of the short-range interaction model are being compensated for by a Monte Carlo procedure (Ref. 7) that includes medium- and long-range interactions. Information about these long-range interactions is deduced from a comparison (such as that in Table 1) between the results obtained from the host-guest technique and from X-ray data on proteins. In addition, the Monte Carlo technique is being modified by inclusion of other empirical information about long-range interactions, e.g., specific electrostatic effects, hydrogen bonds, distribution of polar and nonpolar residues with respect to the center of mass of the protein molecule (Ref. 42), hydrophobic nucleating contacts (Ref. 43), etc.

X. CONCLUSION

Studies of random copolymers, using the host-guest technique, have provided insight into the helix-vs-coil preferences of the naturally occurring amino acids--within the framework of a short-range interaction model. Comparison with similar data from X-ray structures of proteins gives information about the role of long-range interactions in determining which segments of a globular protein will be helical. The use of such data (and similar data for other than helical conformational states) in the manner indicated in section IX should help in circumventing the multiple-minimum problem mentioned in section I, and thus provide a means to determine how polypeptide chains fold to achieve the three-dimensional structures of globular proteins.

Acknowledgment - This work was supported by research grants from the National Institute of Arthritis and Metabolic Diseases, National Institutes of Health, U.S. Public Health Service (AM-08465), and from the National Science Foundation (PCM75-08691).

I am indebted to J. S. Anderson, Y. Isogai, R. R. Matheson, F. R. Maxfield, G. Némethy, S. P. Powers, and R. K. Scheule for helpful discussion and comments on this manuscript.

REFERENCES

1. H.A. Scheraga, *Pure and Applied Chem.* 36, 1-8 (1973).
2. A.W. Burgess, P.K. Ponnuswamy and H.A. Scheraga, *Israel J. Chem.* 12, 239-286 (1974).
3. P.Y. Chou and G.D. Fasman, *Biochemistry* 13, (a) 211-222 (1974); (b) 222-245 (1974).
4. B. Robson and R.H. Pain, *Biochem. J.* 141, (a) 869-882 (1974); (b) 883-897 (1974); (c) 899-904 (1974).
5. F.R. Maxfield and H.A. Scheraga, *Biochemistry* 15, 5138-5153 (1976).
6. S. Tanaka and H.A. Scheraga, *Macromolecules* (a) 9, 142-159 (1976); (b) 9, 159-167 (1976); (c) 9, 168-182 (1976); (d) 9, 812-833 (1976); (e) 10, 9-20 (1977); (f) 10, 305-316 (1977).
7. S. Tanaka and H.A. Scheraga, *Proc. Natl. Acad. Sci., U.S.*, (a) 72, 3802-3806 (1975); (b) 74, 1320-1323 (1977).
8. IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry* 9, 3471-3479 (1970).
9. B.H. Zimm and J.K. Bragg, *J. Chem. Phys.* 31, 526-535 (1959).
10. G. Némethy and H.A. Scheraga, *Quart. Rev. Biophys.*, in press.
11. F. Cardinaux, J.C. Howard, G.T. Taylor and H.A. Scheraga, *Biopolymers*, in press.
12. J.C. Howard, F. Cardinaux and H.A. Scheraga, *Biopolymers*, in press.
13. S. Lifson, *Biopolymers* 1, 25-32 (1963).
14. G. Allegra, *J. Polymer Sci. C16*, 2815-2824 (1967).
15. G.W. Lehman and J.P. McTague, *J. Chem. Phys.* 49, 3170-3179 (1968).
16. D. Poland and H.A. Scheraga, *Biopolymers* 7, 887-908 (1969).
17. P.H. von Dreele, D. Poland and H.A. Scheraga, *Macromolecules* 4, 396-407 (1971).
18. P.H. von Dreele, N. Lotan, V.S. Ananthanarayanan, R.H. Andreatta, D. Poland and H.A. Scheraga, *Macromolecules* 4, 408-417 (1971).
19. V.S. Ananthanarayanan, R.H. Andreatta, D. Poland and H.A. Scheraga, *Macromolecules* 4, 417-424 (1971).
20. K.E.B. Platzer, V.S. Ananthanarayanan, R.H. Andreatta and H.A. Scheraga, *Macromolecules* 5, 177-187 (1972).
21. L.J. Hughes, R.H. Andreatta and H.A. Scheraga, *Macromolecules* 5, 187-197 (1972).
22. J.W. van Nispen, D.J. Hill and H.A. Scheraga, *Biopolymers* 16, 1587-1592 (1977).
23. J.E. Alter, G.T. Taylor and H.A. Scheraga, *Macromolecules* 5, 739-746 (1972).
24. H.E. Van Wart, G.T. Taylor and H.A. Scheraga, *Macromolecules* 6, 266-273 (1973).
25. J.E. Alter, R.H. Andreatta, G.T. Taylor and H.A. Scheraga, *Macromolecules* 6, 564-570 (1973).
26. F.R. Maxfield, J.E. Alter, G.T. Taylor and H.A. Scheraga, *Macromolecules* 8, 479-491 (1975).
27. R.K. Scheule, F. Cardinaux, G.T. Taylor and H.A. Scheraga, *Macromolecules* 9, 23-33 (1976).
28. M.K. Dygert, G.T. Taylor, F. Cardinaux and H.A. Scheraga, *Macromolecules* 9, 794-801 (1976).
29. R.R. Matheson, Jr., R.A. Nemenoff, F. Cardinaux and H.A. Scheraga, *Biopolymers* 16, 1567-1585 (1977).
30. D.J.T. Hill, F. Cardinaux and H.A. Scheraga, *Biopolymers*, in press, paper on methionine.
31. Y. Konishi, J.W. van Nispen, G. Davenport and H.A. Scheraga, *Macromolecules*, submitted.
32. Y. Kobayashi, F. Cardinaux, B.O. Zweifel and H.A. Scheraga, *Macromolecules*, submitted.
33. D.J.T. Hill, F. Cardinaux and H.A. Scheraga, *Biopolymers*, in press, paper on distribution of guest residues.
34. M. Gō, F.T. Hesselink, N. Gō and H.A. Scheraga, *Macromolecules* 7, 459-467 (1974).
35. Y. Isogai, G. Némethy and H.A. Scheraga, work in progress.
36. E. Suzuki and B. Robson, *J. Mol. Biol.* 107, 357-367 (1976).
37. S. Tanaka and H.A. Scheraga, work in progress.
38. F.R. Maxfield and H.A. Scheraga, *Macromolecules* 8, 491-493 (1975).
39. A.V. Finkelstein and O.B. Ptitsyn, *Biopolymers* 16, 469-495 (1977).
40. A.V. Finkelstein, O.B. Ptitsyn and S.A. Kozitsyn, *Biopolymers* 16, 497-524 (1977).
41. A.V. Finkelstein, *Biopolymers* 16, 525-529 (1977).
42. S. Rackovsky and H.A. Scheraga, work in progress.
43. R.R. Matheson, Jr., and H.A. Scheraga, work in progress.