

ON THE DOMINANCE OF SHORT-RANGE INTERACTIONS IN POLYPEPTIDES AND PROTEINS

HAROLD A. SCHERAGA

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

ABSTRACT

The development of the concept of the dominance of short-range interactions in polypeptides and proteins is traced. It is shown that certain amino acid residues have a preference to adopt the right-handed α -helical conformation while others exhibit a preference to participate in β -turns. These ideas are being applied in conformational energy calculations to try to determine the native conformations of proteins.

I. INTRODUCTION

In computing protein conformation from amino acid sequence, using empirical energy functions, the problem of the existence of many local minima in the multi-dimensional energy surface is encountered¹. Since (energy) minimization algorithms lead to the nearest local minimum, depending on the starting point, it is desirable to find alternative methods that would lead to initial starting points which might have a reasonable chance of being in the desired potential well; then, energy minimization would lead to the desired local minimum. For reasons outlined in section III, one of the approaches for finding alternative methods led to a consideration of the possible dominance of short-range interactions^{2, 3}. This investigation led to the concept^{2, 3} that the conformation of an amino acid residue in a polypeptide or protein is determined in very large measure (though not exclusively) by the short-range interactions between a sidechain and the atoms of the backbone of the *same* amino acid residue, and is (again in first approximation) essentially independent of interactions with neighbouring sidechains or backbone portions of the chain. This view has recently received further support from a statistical analysis of the conformations of amino acid residues in globular proteins by Finkelstein and Ptitsyn⁴. In this paper, we shall trace the development and application of this concept.

II. DEFINITIONS

As used here, the term 'short-range' refers to an interaction between the sidechain of an amino acid residue with its own backbone. The interaction between the atoms of a given residue with those of any other residue, nearby

in the chain or more remote along the chain (even though, possibly, nearby in space) is termed 'long-range'.

III. THE θ -POINT

The treatment of an ideal homopolymer chain by random-flight statistics leads to the conclusion that some average linear dimension of the chain, e.g. the root-mean-square end-to-end distance, $\langle \bar{r}^2 \rangle^{\frac{1}{2}}$, varies with the square root of the molecular weight⁵. While long-range and excluded volume effects (not included in the random-flight calculation) tend to increase $\langle \bar{r}^2 \rangle^{\frac{1}{2}}$ beyond its *ideal* value, the choice of an appropriate (poor) solvent (in which polymer-polymer contacts are favoured over polymer-solvent contacts) can reduce $\langle \bar{r}^2 \rangle^{\frac{1}{2}}$ to its *ideal* value⁵. Under these conditions (i.e. at the θ -point), the polymer-polymer and polymer-solvent interactions compensate the long-range and excluded volume effects, and the ideal value of $\langle \bar{r}^2 \rangle^{\frac{1}{2}}$ which results is determined by short-range interactions⁵. Although a protein in aqueous salt solution may not be at the θ -point, the possibility existed that its conformation, while not determined exclusively by short-range interactions, might nevertheless be dominated by them. As will be shown here, the dominance of short-range interactions has been demonstrated for the formation of α -helical and non-helical portions of proteins²⁻⁴ and for the formation of β -turns⁶; it remains to be seen whether β -structures also are determined in large measure by short-range interactions[†].

IV. CONFORMATIONAL PREFERENCES WITHIN A SINGLE PEPTIDE UNIT

To examine the validity of the hypothesis that short-range interactions are dominant, a study was made² of the role of these interactions in helix formation for proteins of known structure. In particular, calculations were carried out to obtain the energy of interaction of individual sidechains in lysozyme with sidechains that are nearest neighbours along the backbone, as well as with the backbone groups themselves. It was found that, for various initial backbone conformations (viz. the right- and left-handed α -helices, α_R and α_L , respectively, and the antiparallel pleated sheet structure, β), the conformation of lowest energy after minimization was the same in most cases for a given amino acid residue and was independent of the nature of the next amino acid in the chain. Furthermore, the backbone structures corresponding to the lowest energy (i.e. α_R , β or α_L) showed a high degree of correlation with the so-called helix-making or helix-breaking character of a residue, as determined by earlier *empirical* studies on the identification of α -helical regions in proteins⁷⁻¹⁰. In other words, it appears that the short-range interactions within a given peptide unit may be the physical origin of the so-called helical potential of a residue. In addition, since the sidechain-

† Even though a parallel or anti-parallel β -structure involves hydrogen bonds (which are long-range interactions), nevertheless the possibility exists that certain residues have a preference for β -conformations because of short-range interactions, and the hydrogen bonds might then contribute added (long-range) stability.

SHORT-RANGE INTERACTIONS

sidechain interaction does not usually play a major role in determining conformation, the cooperativity among residues, which is necessary for the formation of a helical segment, may simply be the additive effect of placing some sequence of helix-making residues in a particular region. This suggested a model for helix formation in which each type of peptide unit in proteins of known amino acid sequence was assigned a designation *h* or *c* (helix-making or helix breaking, respectively), based on a study of the energy surface of the peptide unit. Then, from an examination of the *h* or *c* assignments for lysozyme, myoglobin, α -chymotrypsin and ribonuclease, empirical rules were formulated to distinguish between helical and non-helical regions. These rules are: (a) an α -helical segment will be nucleated when at least four *h* residues in a row appear in the amino acid sequence and (b) this helical segment will continue growing toward the C-terminus of the protein until two *c* residues in a row occur, a condition that terminates the helical segment. With these rules, it was possible to predict the helical or non-helical state of 78 per cent of the residues of the four proteins mentioned above³.

With the later availability of the x-ray structures of seven proteins, the validity of these rules was examined further¹¹. It was observed that, if a dipeptide ever occurred at the C-terminus of a helical region, it had a low probability of occurring elsewhere in a helical region and as high as a 90 per cent probability of occurring elsewhere in non-helical regions. It was also found that those residues designated as *c* tended to predominate at the C-termini of helical segments. These results constitute an experimental demonstration of the validity of rule *b* above. Finkelstein and Ptitsyn⁴ also made a statistical analysis of the conformations of amino acid residues in proteins of known structure, and came to similar conclusions, viz. that short-range interactions are dominant, in that single residues can be classified as helix-making or helix-breaking and that sidechain-sidechain interactions play a minor role in determining the conformational preference of a given amino acid residue.

At this point, it is of interest to consider the factors which determine the conformational preference of a given amino acid residue. The conformational entropy of a residue in the random coil state must be overcome by favourable energetic factors in order for the residue to be helix-making; otherwise, it will be helix-breaking. Glycyl residues, with no sidechains, have no favourable energetic factors to enhance helix formation; thus, the entropy of the coil makes glycyl residues helix-breaking¹². When a β -CH₂ group is added, the resulting non-bonded interactions tend to favour the α_R conformation^{2, 12}. However, as in Asn which is helix-breaking, electrostatic interaction between a polar sidechain group and the polar backbone amide group de-stabilizes the α_R conformation relative to other conformations. In Gln and Glu, the electrostatic effect is weaker because of the greater distance between the backbone amide group and the polar sidechain group (resulting from entropically-favoured extended sidechain conformations); hence, the preferred conformation for Gln and Glu is α_R . Recently, an extensive series of conformational energy calculations (including the computation of statistical weights) was carried out for the *N*-acetyl-*N'*-methyl amides of all twenty naturally-occurring amino acids¹³. From these calculations, it is possible to assess how the various energetic factors contribute to the con-

formational preferences of each residue. For example, the sidechains of both Ser and Asp can form hydrogen bonds with the nearby backbone amide groups when these residues are in non-helical conformations; thus, Ser and Asp are helix-breaking.

V. QUANTITATIVE SPECIFICATION OF HELIX-MAKING AND HELIX-BREAKING CHARACTER

Having demonstrated that the conformation of an amino acid residue in a protein is determined largely by short-range interactions, and, thus, in first approximation is essentially independent of the chemical nature of its neighbours, it becomes desirable to have a quantitative scale to specify the helix-making and helix-breaking character of the twenty naturally-occurring amino acids—instead of the earlier^{2, 3} assignment of all amino acids to two categories, *h* or *c*. A model which suggests itself is the helix-coil transition in homopolymers; i.e. the Zimm-Bragg parameters σ and s (cf. ref. 14), which characterize the transition curve, would appear to provide a quantitative basis for specifying the helix-making and breaking tendency of any amino acid in its corresponding homopolymer (and, therefore, in a protein, since short-range interactions dominate in both cases). Because of certain experimental problems discussed elsewhere^{1, 15, 16}, homopolymers cannot be used for this purpose, and resort is had instead to the use of random copolymers of two components—a helical host, for which σ and s are known, and the guest residues; from the effect of increasing amounts of the guest residues on the helix-coil transition curve of the homopolymer of the host residues, it is possible to determine σ and s for the guest residues^{1, 15, 16}. Thus far, these experiments have been carried out for the following guest residues: Gly¹⁷, Ala¹⁸, Ser¹⁹ and Leu²⁰, and the results are shown in Table 1. It can be seen that Gly and Ser are helix-breakers, Gly more so than Ser (because $s < 1$), and Ala and Leu are helix-makers, Leu more so than Ala. Since the experiments were carried out in aqueous solution¹⁶⁻²⁰, the resulting experimental values of σ and s contain all energetic and entropic contributions (including solvation) which determine the conformational preference.

Table 1. Experimental values of σ and s determined by the host-guest technique

Temp., °C	s			
	Glycine ^a (ref. 17)	L-Alanine ^b (ref. 18)	L-Serine ^c (ref. 19)	L-Leucine ^d (ref. 20)
0	0.51	1.08	0.73	1.10
10	0.55	1.08	0.77	1.12
20	0.59	1.07	0.78	1.14
30	0.62	1.06	0.79	1.14
40	0.63	1.04	0.79	1.13
50	0.63	1.02	0.78	1.11
60	0.63	1.01	0.74	1.09
70	0.61	1.00	0.72	1.06

^a $\sigma = 1 \times 10^{-3}$; ^b $\sigma = 8 \times 10^{-4}$; ^c $\sigma = 7.5 \times 10^{-3}$; ^d $\sigma = 33 \times 10^{-4}$.

VI. HELIX PROBABILITY PROFILES

The experimental values of σ and s can be used to obtain information about the conformation of any specific sequence of amino acids, e.g. that of a protein. However, since the values of σ and s were obtained from the Zimm–Bragg theory, which is based on the one-dimensional Ising model, we cannot treat the *native* protein molecule since its conformation is, in some measure, influenced by long-range interactions which are not taken into account in the Zimm–Bragg theory. Since the *denatured* protein is devoid of tertiary structure and hence, presumably, of long-range interactions other than excluded volume effects, the polypeptide conforms to the one-dimensional Ising model. Thus, above the denaturation temperature, we may apply the Zimm–Bragg formulation to this copolymer of ~ 20 amino acids to determine the probability that any given residue of the chain will be in the α_R or in the random coil conformation, respectively²¹. It will then be shown that there is a correlation between the calculated α_R probability profile of the *denatured* protein and the experimentally observed α_R regions in the corresponding *native* structures; i.e. in many cases, those regions in the denatured protein which exhibit a propensity for being in the α_R conformation correspond to the α_R regions observed in the native protein.

The partition function Z , and the probability, $P_H(i)$, that the i th amino acid (of type A) in a chain of N residues is in the α_R conformation are given by

$$Z = (0, 1) \left[\prod_{j=1}^N \mathbf{W}_A(j) \right] \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (1)$$

and

$$P_H(i) = (0, 1) \left[\prod_{j=1}^{i-1} \mathbf{W}_A(j) \right] \frac{\partial \mathbf{W}_A(i)}{\partial \ln s_A(i)} \left[\prod_{j=i+1}^N \mathbf{W}_A(j) \right] \begin{pmatrix} 1 \\ 1 \end{pmatrix} / Z \quad (2)$$

where $\mathbf{W}_A(j)$ is the matrix of statistical weights for the j th residue which is of amino acid type A, viz.

$$\mathbf{W}_A(j) = \begin{pmatrix} s_A(j) & 1 \\ \sigma_A(j)s_A(j) & 1 \end{pmatrix} \quad (3)$$

$s_A(j)$ is the statistical weight assigned to this residue when it is in an α_R conformation and preceded by a residue in the α_R conformation, and $\sigma_A(j)s_A(j)$ is the statistical weight assigned to this residue when it is in an α_R conformation and preceded by a residue in the random coil conformation. The use of equation 2 to compute $P_H(i)$ automatically includes the cooperativity which is characteristic of the nearest-neighbour one-dimensional Ising model.

Pending the acquisition of data, such as those of *Table 1*, for the remainder of the twenty naturally-occurring amino acids, the set of amino acids has been grouped into three categories (all with σ taken as 5×10^{-4}), viz. helix-breakers (with $s = 0.385$), helix-formers (with $s = 1.05$), and helix-indifferent (with $s = 1.00$). Taking into account the limited data of *Table 1*, the earlier h and c assignments of Kotelchuck and Scheraga^{2, 3}, and the results of an information-theory analysis by Pain and Robson²², the amino acids are assigned as in *Table 2*. It should be emphasized that these tentative values of

HAROLD A. SCHERAGA

Table 2. Assignment of amino acid residues to three categories according to helix-forming power

Helix-breaker		Helix-indifferent			Helix-former		
Gly	Pro	Lys	Asp	Arg	Val	His	Met
Ser	Asn	Tyr	Thr	Cys	Gln	Ala	Leu
				Phe	Ile	Trp	Glu

σ and s are used here only pending completion of experiments which will complete Table 1 for the remaining amino acids.

Helix probability profiles for eleven proteins have been calculated from equation 2, using the values of σ and s discussed above and the assignments of Table 2²³‡. From these curves it appears that there is a close correlation between the propensity of a particular amino acid residue to be in the α_R conformation in the denatured protein and its occurrence in a helical region in the globular structure of the corresponding native protein. On this basis, it was suggested²¹ that, during renaturation, the protein chain acquires *specific* long-range interactions which stabilize the helical regions which tend to form in certain portions of the chain; i.e. folding of the polypeptide chain into the native conformation of a protein is thought to occur by incipient formation of α -helical or other ordered structural regions (among those residues with a propensity to be helical) stabilized by specific long-range interactions, with the remainder of the protein molecule then folding around these stabilized helical regions.

Consistent with this view, it is found²³ that, despite amino acid substitutions in a series of 27 species of cytochrome c proteins, there is a striking similarity in their helix probability profiles, and a good correlation with the location of the helical regions in the (x-ray determined) structure of the horse and bonito proteins. It appears that amino acid substitutions may be tolerated in evolution, provided that the helix-forming or helix-breaking tendency (i.e. values of σ and s) of each amino acid residue is preserved, thereby enabling the altered protein to maintain the same three-dimensional conformation and, hence, the same biological function.

Application of this approach to lysozyme and α -lactalbumin²⁵, two different proteins with striking homologies in their amino acid sequence, led to very similar helix probability profiles. This result supports earlier suggestions²⁶⁻²⁸ that the two proteins might have similar three-dimensional structures, and, if so, again demonstrates the conservative nature of amino acid replacements (as far as helix-forming power in homologous proteins is concerned) which was found for the cytochrome c proteins.

VII. β -TURNS

While the protein chain must fold in order to enable remote helical or other ordered structural regions to approach each other to be stabilized by long-range interactions, it is felt that these long-range interactions are not

‡ More detailed information about the conformational state of each residue is provided by a recently-developed eight-state model for the helix-coil transition in homopolymers and specific-sequence copolymers^{12, 24}.

SHORT-RANGE INTERACTIONS

brought into play by a *chance* encounter of the ordered regions⁶. Instead, there is a tendency for bends or β -turns to occur among certain amino acid residues, thereby 'directing' the encounter of the ordered regions. From a statistical analysis of the amino acid compositions of the bends in three proteins, it has been possible to formulate rules for the existence of β -turns in general. Application of these rules to several other proteins led to a high degree of correlation between the predicted regions where the β -turns should appear, and their existence in the (x-ray determined) structure⁶. A discussion of β -turns in proteins has also been presented by Kuntz²⁹. It is of interest that residues like Gly, Ser and Asp, which have a low tendency toward helix formation, have a high propensity to form β -turns.

We have just completed a study³⁰ of the bends found in the native structures of eight proteins. The 135 bends which were located could be grouped among ten types, and over 40 per cent of the bends did not possess a hydrogen bond between the C=O of residue i and the NH of residue $i + 3$. In addition, conformational energy calculations were carried out on three pentapeptides with amino acid sequences found as bends in the native structure of α -chymotrypsin. The results indicate that the bends occur not only in the whole molecule, but also in the pentapeptide; i.e. the observed bends were the conformations of lowest energy even in the pentapeptides. The stability of the bends, compared to those of other structures, arises principally from side-chain-backbone interactions (e.g. a hydrogen bond between the sidechain COO⁻ of Asp in position $i + 3$ and the backbone NH of the residue in position i) rather than from i to $i + 3$ backbone-backbone hydrogen bonds. This result is consistent with the observation⁶ that residues with small polar sidechains, such as Ser, Thr, Asp and Asn, are found frequently in bends, presumably because these residues can interact most strongly with their immediate backbones.

From the above discussion, the following picture of the folding of the polypeptide chain emerges: helical (or other ordered structural) regions tend to form in certain regions of the amino acid sequence of the polypeptide chain, in response to short-range interactions. These are stabilized, however, only when long-range interactions come into play. This is brought about by the formation of β -turns among *specific* amino acid residues, also on the basis of short-range interactions, thereby enabling the ordered regions to approach each other. The remainder of the polypeptide chain then folds around these one or more regions of interacting ordered structures.

VIII. APPLICATION OF CONCEPT OF DOMINANCE OF SHORT-RANGE INTERACTIONS

It is possible that rough models of the three-dimensional structure of a protein can be formulated by applying the above ideas to predict the regions of the chain where helical regions and β -turns occur. It remains to be seen whether similar ideas can be used to predict where parallel and anti-parallel pleated sheet conformations arise. If such a rough model can be obtained, it should be possible to refine it by conventional energy-minimization procedures¹ without encountering the multiple-minima problem.

These ideas are presently being tested³¹ by conformational energy cal-

culations on trimers, pentamers, heptamers and nonamers from lysozyme. The conformation of the central residue of each of these oligopeptides is varied while that of the remainder of the peptide is maintained in the (x-ray determined) observed structure. Preliminary results indicate that the correct conformation of the central residue can be obtained (as that which minimizes the energy of the oligopeptide) when not only the short-range interactions in the central residue are included, but also longer and longer range interactions as the length of the oligopeptide is increased.

NOTE ADDED IN PROOF

1. *Table 1* has been extended to include L-phenylalanine³² and L-valine³³.
2. It appears that extended (β) structures are also determined, in large measure, by short- and medium-range interactions^{31,34}.

REFERENCES

- 1 H. A. Scheraga, *Chem. Revs.*, **71**, 195 (1971).
- 2 D. Kotelchuck and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **61**, 1163 (1968).
- 3 D. Kotelchuck and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **62**, 14 (1969).
- 4 A. V. Finkelstein and O. B. Ptitsyn, *J. Molec. Biol.* **62**, 613 (1971).
- 5 P. J. Flory, *Principles of Polymer Chemistry*, chap. 14, Cornell University Press: Ithaca, N.Y. (1953).
- 6 P. N. Lewis, F. A. Momany and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **68**, 2293 (1971).
- 7 A. V. Guzzo, *Biophys. J.* **5**, 809 (1965).
- 8 B. H. Havsteen, *J. Theoret. Biol.* **10**, 1 (1966).
- 9 J. W. Prothero, *Biophys. J.* **6**, 367 (1966).
- 10 M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967).
- 11 D. Kotelchuck, M. Dygert and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **63**, 615 (1969).
- 12 M. Gö, N. Gö and H. A. Scheraga, *J. Chem. Phys.* **54**, 4489 (1971).
- 13 P. N. Lewis, F. A. Momany and H. A. Scheraga, *Israel J. Chem.* in press.
- 14 B. H. Zimm and J. K. Bragg, *J. Chem. Phys.* **31**, 526 (1959).
- 15 P. H. Von Dreele, D. Poland and H. A. Scheraga, *Macromolecules*, **4**, 396 (1971).
- 16 P. H. Von Dreele, N. Lotan, V. S. Ananthanarayanan, R. H. Andreatta, D. Poland and H. A. Scheraga, *Macromolecules*, **4**, 408 (1971).
- 17 V. S. Ananthanarayanan, R. H. Andreatta, D. Poland and H. A. Scheraga, *Macromolecules*, **4**, 417 (1971).
- 18 K. E. B. Platzer, V. S. Ananthanarayanan, R. H. Andreatta and H. A. Scheraga, *Macromolecules*, **5**, 177 (1972).
- 19 L. J. Hughes, R. H. Andreatta and H. A. Scheraga, *Macromolecules*, **5**, 187 (1972).
- 20 J. Alter, G. T. Taylor and H. A. Scheraga, *Macromolecules*, **5**, 739 (1972).
- 21 P. N. Lewis, N. Gö, M. Gö, D. Kotelchuck and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **65**, 810 (1970).
- 22 R. H. Pain and B. Robson, *Nature, London*, **227**, 62 (1970).
- 23 P. N. Lewis and H. A. Scheraga, *Arch. Biochem. Biophys.* **144**, 576 (1971).
- 24 N. Gö, P. N. Lewis, M. Gö, and H. A. Scheraga, *Macromolecules*, **4**, 692 (1971).
- 25 P. N. Lewis and H. A. Scheraga, *Arch. Biochem. Biophys.* **144**, 584 (1971).
- 26 K. Brew, T. C. Vanaman and R. L. Hill, *J. Biol. Chem.* **242**, 3747 (1967).
- 27 R. L. Hill, K. Brew, T. C. Vanaman, I. P. Trayner and P. Mattock, *Brookhaven Symp. Biol.* **21**, 139 (1968).
- 28 W. J. Browne, A. C. T. North, D. C. Phillips, K. Brew, T. C. Vanaman and R. L. Hill, *J. Molec. Biol.* **42**, 65 (1969).
- 29 I. D. Kuntz, *J. Amer. Chem. Soc.* **94**, 4009 (1972).
- 30 P. N. Lewis, F. A. Momany and H. A. Scheraga, *Biochim. et Biophys. Acta*, **303**, 211 (1973).
- 31 P. K. Ponnuswamy, P. K. Warne and H. A. Scheraga, *Proc. Nat. Acad. Sci., Wash.* **70**, 830 (1973).
- 32 H. E. Van Wart, G. T. Taylor and H. A. Scheraga, *Macromolecules*, **6**, 266 (1973).
- 33 J. Alter, R. H. Andreatta, G. T. Taylor and H. A. Scheraga, *Macromolecules*, in press.
- 34 P. K. Ponnuswamy, A. W. Burgess and H. A. Scheraga, submitted to *Israel J. Chem.*