Analysis of the Code Relating Sequence to Conformation in Globular Proteins

THE DISTRIBUTION OF RESIDUE PAIRS IN TURNS AND KINKS IN THE BACKBONE CHAIN

By BARRY ROBSON and ROGER H. PAIN
Department of Biochemistry, University of Newcastle upon Tyne,
Newcastle upon Tyne NE1 7RU, U.K.

(*Received* 24 *September* 1973)

1. The residue pair is considered as the fundamental unit which differentiates α-helix, β-pleated sheet and the various turns and kink structures of the protein backbone. 2. The HPLG alphabet (Robson & Pain, 1974) is used to group pairs of residues, giving 16 possible conformational pairs, all of which are found with differing frequencies in the nine proteins examined. 3. The frequencies of occurrence of the 16 different types of turn or kink are analysed in relation to the constituent amino acids. Those containing the L or G conformation are of low frequency and are grouped for purposes of this analysis. 4. The distribution of amino acids within all the conformational pairs is non-random, with distinct preferences shown by certain residues.'5. All pairs containing an L or G conformation require the presence of a glycine or a proton-donor side chain. 6. The results are discussed in terms of the determination of these 'random' structures by local interactions.

The application of information-theory analysis to breaking the code relating sequence to conformation in globular proteins has broadened our understanding of the influence that an amino acid residue has on its own conformation and on that of its neighbours. Perhaps the most important generalization that has been established is that the formation of certain backbone structures is largely determined by the character of the residues in the immediate neighbourhood of the structure and not by distant, or tertiary, interactions. The quality of predictions obtained for the location of helices in globular proteins on the basis of local information derived from residues only ±8 in number from the residue whose conformation is being studied, has shown that this structure is determined by local interactions between side chains and backbone (Robson & Pain, 1971). The relative success of a variety of other statistical approaches in predicting helices reinforces this general conclusion (for a review see Robson, 1972).

More recently, with an expansion in the number of protein conformations revealed by X-ray crystallography, enough data has become available to enable other types of structure within globular proteins to be analysed. Nagano (1973) has used a statistical analysis to predict the location of helices, loops and β-structures. An empirical prediction method has shown that ϕ, ψ backbone angles may, under circumstances, be predicted from an examination of residues taken in triplets (Wu & Kabat, 1973). The contribution that single residues alone make to the determination of ϕ, ψ angles has been analysed (Robson & Pain, 1974) by defining a stereochemical alphabet to relate backbone angle to intraresidue information measures. From this analysis it is apparent that β-pleated-sheet structures are, to a considerable extent, the result of interactions between the residue side chain and its own backbone.

The question arises as to whether all other, less regular, structures are dictated by local interactions or not. Intuitively, it has been assumed that these conformations, frequently termed 'random-coil' structures, are determined mainly by the tertiary folding of locally determined regular structures. Recently, considerable interest has been aroused in the possible importance of those regions of the globular protein that are involved in an approx. 180° reversal of the backbone. Such structures were first expected to be especially stable on the basis of an analysis of those conformations that were both sterically permissible and which formed stable hydrogen bonds between the backbone carbonyl group of the jth residue and the backbone amide group of the $(j+4)$th residue (Venkatachalam, 1968). Such structures and their variants are found in globular proteins and are often referred to as β-turns because they appear at the hair-pin bend where an approximately extended chain turns back on itself to form an anti-parallel β-pleated sheet. However, they are also found very frequently elsewhere, and are reported to make up a great deal of the backbone conformation of globular proteins (Crawford *et al.*, 1973). They are more generally referred to as reverse turns. Attempts to predict the occurrence of the reverse turn have met with more limited success (e.g. Lewis *et al.*, 1971; Nagano, 1973), and this reflects a relative lack of knowledge concerning the formation of turns which are quite unlike those of the α-helix or β-pleated sheet.

The recently developed stereochemical alphabet (Robson & Pain, 1974) lends itself to an investigation of all types of turns. An unambiguous definition of reverse turns enables the abundance of the different types of turn to be analysed in relation to the amino acid type of the constituent residues. The approach is wholly empirical, making no assumptions *a priori* about the physical interactions that are likely to lead to stable backbone conformations, and leads to the conclusion that local interactions play an important part in the determination of the backbone angles of all the so-called random-coil regions of globular proteins. This is shown to apply not only to turns possessing an internal hydrogen-bonded bridge, but also to that large proportion of turns whose hydrogen-bonding groups are engaged outside the turn.

Methods

The backbone conformation of a reverse turn is defined by the backbone angles ϕ_j, ψ_j, ϕ_{j+1}, ψ_{j+1} which belong to the *j*th residue and its successor. With these angles specified and the peptide-bond ω assumed planar and *trans* (note, however, the special case of proline; see, for example, Lewis *et al.*, 1971), all the backbone atoms between and including the C_α atoms of the ($j-1$)th and ($j+2$)th residues can unambiguously be located in space with respect to one another.

The definition of a reverse turn proposed by Crawford *et al.* (1973) involves four amino acids. Although consideration of these four residues is relevant to prediction of reverse turns, only the inner two residues contain the bonds whose orientations actually define a reverse turn. Hence, the number of residues defined by Crawford *et al.* (1973) as being involved in reverse turns could be up to twice the number of residues actually in the reverse-turn conformation.

Reverse turns and related structures can now be studied by analysis of the conformation of a pair of residues, which may be considered as the fundamental units required to differentiate α-helix, β-pleated sheet and the various turn structures.

The original ϕ, ψ data were obtained from nine proteins of known sequence and conformation, as listed by Robson & Pain (1974). Conformations were defined not explicitly as ϕ, ψ angles, but by the general area of the ϕ, ψ diagram to which those angles belong. These areas are the domains H, P, L and G defined on Fig. 1, which make use of the natural distribution of conformations as shown by Robson & Pain (1974). Thus, any sequence of angles can be written down by using the stereochemical alphabet of letters H, P, L and G. For example, a sequence of conformation (21°, 106°), (−67°, −42°), (104°, 175°) can be written as LHG. Most of the different types of reverse turn can be distinguished in the data by this method. However, the closely related reverse folds I and III of Venkatachalam (1968), which differ only by a 30° distortion of each bond of the second residue in the pair, are both classified as HH turns in our nomenclature. This is also true of the mirror images of these backbone conformations I′ and III′ which we call LL.

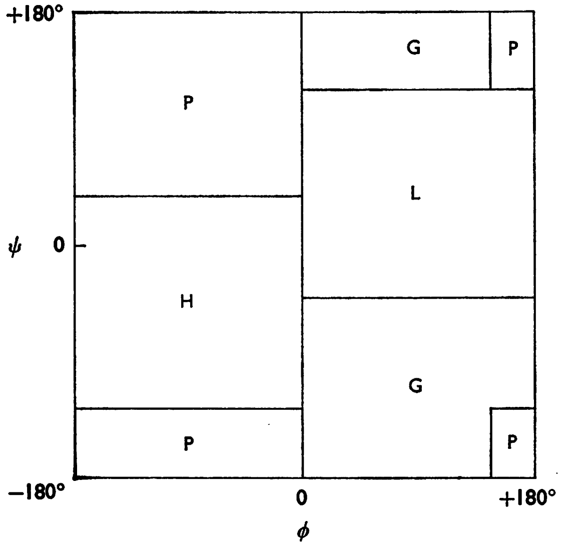

Fig. 1. *Domains, derived from the natural distribution of amino acid-residue conformations in globular proteins, which are used as the basis of the HPLG alphabet* (*cf.* *Robson & Pain,* 1974)

The alphabet into which the conformations of the proteins have been translated and the types of reverse turn described are intentionally more general than that used by other authors (Crawford *et al.*, 1973; Lewis *et al.*, 1973) and allow for those perturbations of the ideal Venkatachalam (1968) conformations which occur so frequently in proteins. In particular, the sample of HH pairs also encompasses all those pairs drawn from regions that are usually considered as being α-helical. This is by no means a purely artificial effect of our broader set of criteria for defining turns, for even in the analysis of Crawford *et al.* (1973), which uses narrower criteria, over one-third of type I plus type III turns belong to α-helical regions. In other words, we make no attempt here to distinguish between a distorted type I turn, a distorted type III turn and a distorted helical turn. On the other hand, distorted reverse turns, which have been grouped by various degrees of relaxation of the criteria used to define a reverse turn [cf. 'near reverse' and 'open' turns of Crawford *et al.* (1973); type IV and VII reverse turns of Lewis *et al.*, 1973], are here defined explicitly in terms of the stereochemical HPLG alphabet.

The conformation of these proteins, written as a stereochemical alphabet, was analysed by recording the numbers of different kinds of pair within it. For example, an imaginary short polypeptide of conformation HHPL would be analysed into pairs HH, HP and PL. The type and order of amino acids in these pairs were also recorded.

Results and Discussion

Data for all the possible pairs that can arise from combinations of H, P, L and G are listed in Table 1, which includes the relation of combinations to those turns already characterized by other workers. It can be seen from column (3) that all possible pairs are in fact observed, though with by no means equal abundance. As might be expected, pairs HH and PP, which are units of right-hand α-helical and pleated-sheet structures respectively, are the most abundant pair conformations, comprising 65% of the sample. It does not follow that all HH and PP pairs make up those structures that are normally classified as helix and pleated sheet, although Table 2 shows that a high proportion do so. Comparison of columns (3) and (4) of Table 1 shows that both HH and PP conformations are statistically co-operative in the sense that the % abundance of these pairs is considerably higher than would be expected on the basis of a random distribution of H, P, L and G. The latter expected abundance is estimated from the product of the abundances of the single residue conformations, e.g. (expected % abundance of HP) = (% abundance of H) × (% abundance of P)/100.

Of principal interest are the % abundances of the less abundant pair conformations. These include the reverse turns II, I′, II′, III′ of Venkatachalam (1968) and bends V and V′ of Lewis *et al.* (1973). They also include conformation pairs that do not constitute reverse turns or V or V′ bends, yet which make up 29% of the sample. Since, with the exception of PP pairs which include the extended chain conformation, all the pairs represent some kind of turn or 'kink' in the polypeptide backbone, there are clearly many kinds of turn that have yet to be investigated. In fact, with the exclusion of HH turns the remaining reverse turns (namely II, II′, V, and V′) make up only 7% of the sample.

It is important to know whether reverse turns are determined, at least in part, by the amino acid side chains of the backbone forming the turn. The possibility that they are is borne out by the analysis of Crawford *et al.* (1973), which indicates a fairly strong correlation between the existence or otherwise of turns and the amino acid sequence. It is also possible that the local amino acid sequence is responsible for the formation of the other kinds of turn which have not yet been investigated, namely HP, HL, HG, PH, LH, LP, LG, GL and GG. HH and PP turns are the most abundant and HP and PH turns are very frequent despite the fact that their expected % abundance [Table 1, column (4)] is decreased by the co-operativity of HH and PP structures. There is therefore sufficient data to carry out an analysis of the amino acid content and position within HH, PP, HP and PH pairs, and the results are presented in Tables 3 and 4. The percentage distribution between a given position in residue pair conformations is tabulated

Table 1. *Analysis of possible conformations of residue pairs*

Note that all these pairs, with the exception of PP, would necessitate a kink in an otherwise extended chain.

(1) Pair	(2) Convention used by other authors for corresponding conformations	(3) % of total pairs	(4) % of total pairs expected by chance§
HH	1 and III†	33.6	21.1
HP	*	9.6	20.2
HL	*	2.3	2.7
HG	*	1.2	1.3
PH	*	9.4	20.2
PP	*	31.2	19.4
PL	II†	2.9	2.6
PG	V‡	1.0	1.2
LH	*	1.7	2.7
LP	*	3.4	2.6
LL	I′ and III′†	0.6	0.4
LG	*	0.3	0.2
GH	II′†	1.7	1.3
GP	V′‡	0.4	1.2
GL	*	0.4	0.2
GG	*	0.4	0.1

* Not reverse turns.

† From Venkatachalam (1968).

‡ From Lewis *et al.* (1973).

§ See the text. Percentage abundances %P = 44%; %L = 6%; %G = 3%; %H = 46%.

Table 2. *Occurrence of H, P, L or G conformations in runs*

	Conformation pairs (% of all pairs)	Conformation pairs in runs of three or more residues in the same conformation (% of all pairs)
HH	33.6	30.7
PP	31.2	28.1
LL	0.6	0.0
GG	0.4	0.0

Table 3. *Percentage distribution of each amino acid between residue pair conformations*

The frequency of occurrence of a residue as one member (i.e. first or second) of a pair is expressed as a percentage of its occurrence as that member of all pairs.

Residue	H	H	P	P	H	P	P	H	L/G	X	X	L/G	Percentage of residue in L/G turns
Gly	13	11	20	13	5	9	4	3	50	16	8	47	54
Ala	52	49	26	28	5	6	5	10	3	5	9	2	9.4
Val	35	28	45	45	6	6	6	9	2	10	6	2	7.6
Leu	36	41	37	43	7	1	7	6	3	7	9	3	9.0
Ile	41	35	38	33	11	7	7	11	3	11	1	3	6.9
Ser	24	28	25	31	17	12	18	12	7	11	9	6	14.5
Thr	25	26	37	38	14	8	10	15	5	8	9	5	11.9
Asp	32	34	22	20	12	17	17	10	5	14	13	5	14.7
Glu	60	59	22	16	1	10	8	4	3	10	6	1	8.3
Asn	24	31	14	25	13	16	25	6	14	6	10	16	20.1
Gln	29	31	44	35	10	11	8	9	4	11	6	4	8.7
Lys	42	39	25	24	9	10	7	12	7	7	10	7	12.0
His	33	42	20	26	11	9	11	12	9	2	16	9	15.2
Arg	32	29	34	31	12	12	9	12	5	10	9	5	12.2
Phe	35	33	41	37	8	8	6	6	6	10	4	6	10.5
Tyr	19	24	40	42	10	8	7	8	10	8	15	10	17.6
Trp	39	48	27	10	15	29	12	13	3	0	3	0	1.5
Cys	24	36	50	38	12	14	2	5	2	5	10	2	9.5
Met	53	52	37	26	0	17	5	0	0	4	5	0	4.7
Pro	31	8	37	52	10	4	13	24	0	10	8	0	8.2

Table 4. *Percentage occupancy by amino acids of each residue pair conformation*

Residue	H	H	P	P	H	P	P	H	L/G	X	X	L/G
Gly	3	3	6	4	4	8	4	3	48	15	8	46
Ala	13	12	7	8	4	5	4	9	3	4	9	3
Val	7	6	10	10	4	4	4	7	2	8	5	2
Leu	8	9	8	10	6	1	6	4	2	5	7	2
Ile	6	5	6	5	6	4	4	6	1	6	1	2
Ser	6	7	6	8	14	10	15	10	6	10	8	6
Thr	4	6	9	9	11	6	8	12	4	7	8	4
Asp	9	4	3	3	5	7	7	4	2	6	6	2
Glu	4	9	4	3	1	5	4	2	2	5	3	1
Asn	5	5	2	4	7	9	14	4	8	4	6	11
Gln	10	3	5	4	4	4	3	4	2	4	2	2
Lys	3	9	6	6	8	8	6	9	6	6	10	6
His	4	4	2	2	4	3	4	4	3	1	6	3
Arg	3	3	4	4	5	5	4	5	2	4	4	2
Phe	2	3	4	4	3	3	2	2	2	4	2	2
Tyr	3	3	5	6	4	4	3	4	5	4	7	5
Trp	2	3	2	1	4	6	3	3	1	0	1	0
Cys	2	3	5	3	4	4	1	1	1	2	3	1
Met	3	2	2	1	0	3	1	0	0	1	1	0
Pro	3	1	4	6	4	1	4	9	0	4	3	0

(Table 3) for each amino acid. For example, it shows that 52% of alanine residues occurring as the *N*-terminal of all residue pairs are in the H conformation and are followed by a residue in the H conformation. Asparagine has a high frequency of occurrence (25%) as the first member and proline (24%) as the second member in the PH turn.

The distribution of amino acids within given conformations is listed in Table 4. It shows, for example, that 29% of all P positions in the PH turn are occupied by serine and asparagine. There is, as yet, insufficient data for a detailed analysis of the remaining turns HL, HG, LH, LP, LG, GL and GG, which all contain the infrequent conformations L and

Table 5. *Character of residues in conformation pairs containing G or L conformations*

Pair conformation	% of pair conformations without glycine or potential hydrogen-bonding side chains*	% of pair conformations without glycine or potential proton-donor side chains
HH	17	32
HP	10	16
PH	12	17
PP	19	41
Any pair containing G or L	2	5

* Potential hydrogen-bonding side chains are Ser, Thr, Asp, Asn, Glu, Gln, Lys, His, Arg, Tyr, Trp. Asp and Glu are considered as proton acceptors only; Cys is considered as a weak donor only.

G. However, certain preliminary conclusions may be drawn if these data are pooled by defining pairs (L/G)X and X(L/G) where (L/G) is L or G and X may be H, P, L or G. This analysis is also included in Tables 3 and 4.

The disadvantage of pooling the pair conformations is that information about distribution between the individual conformations is lost. However, glycine is so abundant in the pairs containing L or G conformations that its distribution can be investigated. As shown in Tables 3 and 4, glycine exhibits a strong tendency to be the member of the amino acid pair that adopts the L or G conformation. However, it occasionally appears in the H or P conformation and the non-glycine member adopts the L or G conformation. The final column of Table 3 shows the percentage of each residue that occurs in G- or L-containing turns. Again, glycine stands out, with 54% of all glycine being found in this conformation. Further, serine, aspartic acid, asparagine, histidine and tyrosine also occur with a higher frequency in the G or L conformation compared with other residues. In contrast, only one tryptophan in a hundred occurs in this conformation. The difference between glutamic acid and aspartic acid and between their respective amides is marked. Tables 3 and 4 show that there is a definite relation between residue type and the conformation of the turns in which they appear, not only in already well-established cases, such as HH, but also in those turns that have not before been analysed.

The values in Table 5 reveal another kind of statistical event that is abundant. It is the minimal requirement for either glycine or an amino acid containing a proton donor to be present in any pair conformation containing L or G conformations. In other words, the presence of a proton-donor side chain is apparently able to compensate for the lack of glycine in a structure which requires that $0° \geqslant \phi \lesssim +180°$. This minimal requirement holds true in 95% of pairs involving G or L conformations, whereas, were it a requirement for HH and PP pairs, it would hold true in only 68% and 59% of all cases respectively. Table 3 indicates that although turns are very rich in polar amino acids, for example aspartic acid, the presence of acidic side chains or of any other set of side chains with obviously related properties does not seem to be a minimal requirement since, for example, glutamic acid shows a low preference for turns containing the L or G conformation. Hence, if any pair of residues in a protein is observed to contain neither glycine nor a side chain which is a proton donor, it is very unlikely to contain an L or a G conformation and is most likely to belong to the right-hand α-helix or the β-pleated-sheet conformation.

Conclusions

The occurrence of reverse turns in nine proteins and the distribution of amino acids within the pairs of residues involved was quantitatively analysed by using an HPLG alphabet based on the natural distribution of residues on a ϕ, ψ diagram. The point is made that it is the backbone angles of only two residues that define the reverse-turn conformation and the conformation of other turns that occur in what has usually been referred to as the 'random-coil' region of globular proteins. In this analysis most of the possible pair combinations were treated separately, although the remainder involving L and G conformations and of low-frequency occurrence were pooled and analysed with respect to the distribution of amino acids within the conformations.

The initial finding is that there are many examples of turns and kinks in the backbone of a globular protein that are not units of previously defined structures, i.e. α-helix, β-pleated sheet, reverse turns and V or V′ bends. In fact, Table 1 shows that units of such structures occur rarely with significantly higher frequency, and often with significantly lower frequency than the undefined structures which together make up nearly one-third of the sample studied. Of considerable importance is the demonstration that the distribution of amino acids within these latter turns is non-random, with distinct preference shown by certain residues.

Secondly, the characteristic feature of reverse turns, V and V′ bends, is that they involve at least one hydrogen bond internally. However, in the interior of a globular protein units are brought together in space by the tertiary folding of the molecule, making possible hydrogen bonding between quite different parts of the primary sequence. With units on the surface of the protein, hydrogen bonding of the solvent may occur so that most potential hydrogen bonds are realized one way or another (see, for example, Watson, 1969). It therefore seems likely that the only absolute criterion for a stable turn or kink of any backbone conformation from a hydrogen-bonding point of view is that the turns should tend to occur in a conformation that will not sterically block any potential hydrogen bond that may arise in the stably folded protein. It is concluded that the presence of an 'internal' hydrogen-bond bridge is not a necessary criterion for the formation of a stable turn. Further, the distribution of amino acids within these turns indicates that it is local residue interactions that are important in coding for and leading to the formation of these turns. This is a situation that is already fairly well understood in the formation of β-pleated-sheet structures where there is no possibility of local hydrogen bonding (except, of course, between PP units separated by a reverse turn).

The turns or kinks HP, HL, HG, PH, LH, LG, GL, and GG, which have not previously been analysed, have characteristic amino acid composition and distribution which are just as marked as those of the previously characterized structures. The most obvious characteristic is that all units containing an L or G conformation require a glycine residue or a proton-donor side chain, although the latter and, to a much lesser extent, the glycine residue need not necessarily occur in the G or L conformation.

These observations suggest the likelihood that turns, including those previously undefined, may be predictable on the basis of the amino acids in and around them when suitable prediction procedures, such as the information-theory approach (Robson & Pain, 1971), are applied. This would imply that the physical interactions responsible for turns and kinks in the polypeptide can be understood in terms of local energy interactions, in the same way as has been demonstrated for the α-helix conformation.

This work was supported by a grant from the Medical Research Council. The expert assistance of Miss Moira Bain is gratefully acknowledged.

References

Crawford, J. L., Lipscomb, W. N. & Schellman, C. G. (1973) *Proc. Nat. Acad. Sci. U.S.* **70**, 538–542

Lewis, P. N., Momany, F. A. & Scheraga, H. A. (1971) *Proc. Nat. Acad. Sci. U.S.* **68**, 2293–2297

Lewis, P. N., Momany, F. A. & Scheraga, H. A. (1973) *Biochim. Biophys. Acta* **303**, 211–229

Nagano, K. (1973) *J. Mol. Biol.* **75**, 401–420

Robson, B. (1972) *Spec. Per. Rep. Chem. Soc.*: *Amino Acids, Peptides and Proteins* **4**, 224–229

Robson, B. & Pain, R. H. (1971) *J. Mol. Biol.* **58**, 237–259

Robson, B. & Pain, R. H. (1974) *Biochem. J.* **141**, 869–882

Venkatachalam, C. M. (1968) *Biopolymers* **6**, 1425–1436

Watson, H. C. (1969) *Progr. Stereochem.* **4**, 299–333

Wu, T. T. & Kabat, E. A. (1973) *J. Mol. Biol.* **75**, 13–31