Int. J. Peptide Protein Res. **16**, 1980, 1–11

REVERSALS OF POLYPEPTIDE CHAIN IN GLOBULAR PROTEINS

A.S. KOLASKAR, V. RAMABRAHMAM and K.V. SOMAN*

School of Life Sciences, University of Hyderabad, Hyderabad, and
**Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India*

Received 6th July 1979, accepted for publication 16 January 1980

A simple algorithm has been developed to detect β-bends and 'loops'-chain reversals containing five amino acid residues, using only coordinates of C^{α}-atoms from crystal structure data of globular proteins using the above algorithm. Analysis of bends have showed that the total number of bends in each protein (T_B) is linearly related to total number of non-hydrophobic residues in that protein which in turn is related linearly to total number of amino acid residues. Secondly, we found that a large number of consecutive bends occur in each protein which give rise to on an average only three independent residues per turn. Positional preference of amino acid residues in chain reversals is stressed. Consideration of pairs of amino acid residues in positions (i + 1) and (i + 2) of bends seems to provide a more reliable basis for predicting chain reversals in proteins.

Key words: β-bends and "loops"; polypeptide chain reversals; protein data analysis.

Three-dimensional structures of proteins and polypeptides consist of secondary structures, such as helices, β-structures and chain reversals, and the coil part. Out of these, α-helix and β-structures are well-characterized using both theoretical and experimental methods. However, the interest in the study of reversals in polypeptide chains has gained momentum in the last few years. Venkatachalam (1968) was the first to postulate six types of reversals and called them β-bends; these were later extended by Lewis *et al.* (1973). Many workers (prominently Crawford *et al.*, 1973; Lewis *et al.*, 1973; Chou & Fasman, 1977; Rose & Seltzer, 1977; Rose, 1978; and Levitt, 1978) are concentrating their efforts on devising simple but powerful methods to detect reversals in polypeptide chains by using crystal structure data of globular proteins and to throw some light on the properties of these chain reversals. However, no unique method has yet been developed mainly because only four or five amino acid residues are involved in polypeptide chain reversals as against 10–12 residues in most helices and β-structures.

In the present study we have presented a very simple algorithm, somewhat similar to the one presented earlier by Chou & Fasman (1977), to find the reversals in polypeptide chains from crystal structure data. The algorithm presented here also points out the chain reversals, or 'loops', which contain five amino acid residues, in addition to usual bends which contain only four amino acid residues. The crystal structure data of 21 globular proteins have been analyzed: namely, hen egg white lysozyme; subtilisin BPN'; D-glyceraldehyde 3-phosphate dehydrogenase (GPD); horse Met hemoglobin; flavodoxin;

0367–8377/80/060001–11 $02.00/0

oxidized rubredoxin; tosyl α-chymotrypsin; concanavalin A; thermolysin; chymotrypsinogen A; high potential iron protein (HIPIP); dogfish M4-APO lactate dehydrogenase; parvalbumin; cytochrome b5; trypsin inhibitor; carboxypeptidase A; ribonuclease-S; staphylococcal nuclease; sea lamprey hemoglobin; cytochrome C2; myoglobin. The data was obtained from AMSOM and was supplied by Richard Feldman (National Institute of Health, U.S.A.). The results which are discussed in the succeeding sections indicate the existence of a large number of consecutive bends and the number of the 'loops' is also considerable.

In this work the occurrence of an amino acid residue in a particular position of the chain reversal is stressed, as opposed to the probability of it occurring in bends irrespective of its position in bends. From our analysis, discussed in succeeding sections and from the results of the study made by Chou & Fasman (1977), it is quite clear that certain amino acid residues prefer a particular portion in bends. Therefore, it becomes essential to study the effect of primary sequence on bend-forming amino acid residues in the (i + 1)st and (i + 2)nd positions of bends. In order to get this insight, we have analyzed the data on bends, taking into consideration the pairs of amino acid residues occurring in positions (i + 1) and (i + 2). As will be seen from the discussion of the results, this analysis seemed to give a better basis to predict chain reversals in globular proteins.

METHOD

The values of atomic coordinates of C^{α}-atoms in globular proteins obtained from X-ray crystal structure data are more accurate than the values of ϕ and ψ. Therefore, the criterion developed to detect chain reversals using only coordinates of C^{α}-atoms will be more reliable than the one developed using values of ϕ and ψ. With this in mind, only the coordinates of C^{α}-atoms of the 21 globular proteins mentioned above have been used to find out chain reversals. The distances between each C^{α}-atom of the (i)th and (i + 3)rd residue have been calculated. Those residues for which the distance is below 7.0 Å have the possibility of being either in helix or in chain reversals. To exclude those residues which lie in the helical region, the following criterion has been used. For the cases for which the distance $\langle C_i^{\alpha} \ldots . C_{i+3}^{\alpha} \rangle \leqslant 7.0$ Å, the distances $\langle C_{i-1}^{\alpha} \ldots . C_{i+2}^{\alpha} \rangle$ and $\langle C_{i+1}^{\alpha} \ldots . C_{i+4}^{\alpha} \rangle$ were also computed. If both these distances were found to be in the range of $(\langle C_i^{\alpha} \ldots . C_{i+3}^{\alpha} \rangle \pm 1.5$ Å), then only the residues (i) to (i + 3) were assumed to be in the helical region Under ideal conditions in a helix $(\langle C_{i-1}^{\alpha} \ldots . C_{i+2}^{\alpha} \rangle = \langle C_i^{\alpha} \ldots . C_{i+3}^{\alpha} \rangle = \langle C_{i+1}^{\alpha} \ldots . C_{i+4}^{\alpha} \rangle)$. However, inaccuracies in the coordinates of C^{α}-atoms and distortions in the helices will not give such a condition and so from trial and error the factor of 1.5 Å has been considered a good estimate; our studies on computer graphics have shown that 1.5 Å is a good limit. To find out whether these chain reversals contain four or five residues, the following criterion has been used. If the residues (i) to (i + 3) were found to be in chain reversals, then the distance between C^{α}-atoms of the (i)th and 1 + 4)th residue was calculated and if distance $\langle C_i^{\alpha} \ldots . C_{i+4}^{\alpha} \rangle \leqslant \langle C_i^{\alpha} \ldots . C_{i+3}^{\alpha} \rangle$ then the five residues (i) to (i + 4) were considered to have formed a 'loop'. A simple computer program has been developed to determine all bends and 'loops' in each protein using the above mentioned criteria. The data obtained using the algorithm have been used to calculate total number of amino acid pairs formed by residues in the (i + 1)st and (i + 2)nd, β-bend-forming positions. Using a simple computer program the total number of pairs formed by consecutive amino acid residues in each protein were obtained excluding the first and last two amino acid residues, were classified into 400 types of possible pairs and the number of occurrences of each type of pair in the above mentioned proteins was calculated. Then the frequency of occurrence of each of these pairs in bend-forming positions (i + 1) and (i + 2) was obtained.

RESULTS AND DISCUSSION

In total, 609 bends, including 64 'loops', were detected using the algorithm described above in 21 globular proteins. To save space we have given only the list of 'loops' in each of the proteins in Table 1. A list of bends is available from the authors on request. Lewis *et al.* (1973) have classified bends into 11 categories

TABLE 1

Loops containing five amino acid residues in proteins

Residue No.	Name	(ϕ_{i+1}, ψ_{i+1})	(ϕ_{i+2}, ψ_{i+2})	(ϕ_{i+3}, ψ_{i+3})	Protein
46–50	NTNGS	− 80, − 17	−100, 10	90, − 19	
59–63	NSRWW	− 70, − 10	− 70, − 30	−110, − 40	Lysozyme
60–64	SRWWC	− 70, − 30	−110, − 40	−100, − 30	
17–21	HSQGY	− 79, 7	− 99, − 15	111, − 19	
70–74	GTVAA	− 79, − 34	− 50, − 61	−165, 56	
97–101	GDAGS	− 38, 145	83, − 26	123, 31	Subtilisin BPN'
143–147	VASGV	− 73, − 51	− 65, − 44	137, 36	
259–263	DSFYY	− 30, − 43	− 78, − 19	− 98, − 35	
47–51	DSTHG	− 39, − 40	− 70, − 61	− 82, − 18	
101–105	TIELA	− 49, − 49	− 42, − 81	− 50, − 42	
102–106	IELAS	− 42, − 81	− 50, − 42	− 56, − 76	GPD
129–133	VCGVN	− 71, − 29	− 47, −104	− 74, 0	
198–202	GAAQN	124, − 94	− 79, − 29	− 79, − 37	
312–316	DNEFG	− 64, −100	− 10, − 97	− 49, − 65	
24–28	IESGK	− 70, − 22	− 73, − 14	98, 20	
56–60	MGDEV	− 1, 71	66, 63	57, 68	Flavodoxin
62–66	EESEF	− 65, − 24	− 84, − 36	−118, − 25	
104–108	DGYGC	− 61, − 22	− 88, − 15	99, 18	
35–39	DKTGF	−120, 2	− 75, 2	69, 44	
95–99	NSLTI	− 62, − 8	−114, − 33	− 77, − 51	α-Chymotrypsin
169–173	KKYWG	− 54, − 28	−104, − 54	− 54, − 94	
191–195	CMGDS	− 69, 133	110, − 27	− 76, − 31	
15–19	TDIGD	− 49, − 15	− 99, − 53	150, 11	
97–101	TGLYK	−158, − 75	−110, − 38	− 88, 144	
117–121	SNSTH	− 86, − 41	− 31, − 63	−160, − 90	Concanavalin-A
160–164	SSNGS	− 77, − 12	−114, − 11	132, − 91	
30–34	DAVVL	− 1, −120	− 73, − 3	− 99, − 41	
102–107	QQEGE	− 81, − 11	149, − 42	− 87, 177	
183–187	HSCLV	2, − 58	− 40, − 74	− 65, − 55	
195–199	HGDSV	122, −151	− 51, 115	106, 11	M4-Apo-LDH
207–211	WDAKL	−150, 140	91, −103	180, 29	
241–245	VDVLT	− 43, − 30	− 21, − 52	− 95, − 54	
277–281	DFYGI	−108, −167	14, −139	−102, − 4	
12–16	GVLGD	− 59, − 36	− 87, 7	66, 35	
24–28	YSTYY	−131, 79	61, − 41	−110, 170	
34–38	TRGDG	−122, 119	37, −119	− 79, 7	
115–119	WNGSE	−115, 20	88, 35	159, − 2	
150–154	DYTAG	93, 6	−115, −107	− 66, − 50	Thermolysin
158–162	QNESG	71, −148	− 61, − 44	− 59, − 38	
204–208	SMSDP	−100, − 34	− 76, − 38	−138, 73	
224–228	TQDNG	− 47, 129	34, 58	44, 52	
249–253	THVGV	−126, 102	72, 46	58, 40	
250–254	HYGVS	72, 46	58, 40	−154, 84	

TABLE 1 (continued)

Residue No.	Name	(ϕ_{i+1}, ψ_{i+1})	(ϕ_{i+2}, ψ_{i+2})	(ϕ_{i+3}, ψ_{i+3})	Protein
17–21	NSKST	− 70, − 8	− 78, − 53	−147, 111	Cytochrome b_5
24–28	ILHYK	−155, 116	35, 57	69, 4	
29–33	HPELV	− 60, − 10	−117, 23	−154, 25	
89–93	NYGQN	− 91, 129	95, − 35	− 92, 43	Carboxypeptidase A
65–69	CKNGG	− 35, − 42	− 59, 6	93, 9	Ribonuclease-S
26–30	MYKGQ	−121, 103	64, 30	77, 28	
46–50	HPKKG	− 80, − 78	61, − 2	176, − 76	Staphylococcal nuclease
83–87	DKYGR	− 83, − 13	− 91, 0	82, 26	
93–97	YADGK	−124, 103	68, 34	115, − 23	
14–18	CLACH	− 74, 140	162, − 33	−117, − 34	
15–19	LACHT	162, − 33	−117, − 34	−120, 176	
26–30	DKVGP	50, 109	66, − 89	−100, 158	Cytochrome C_2
104–108	VIAYL	− 50, − 41	− 55, − 93	− 67, − 38	
106–110	AYLKT	− 67, − 38	− 31, − 4	−162, − 30	
107–111	YLKTL	− 31, − 4	−162, − 30	− 83, − 46	
6–10	CTICG				
20–24	TEDGV				Rubredoxin
39–43	CPLCG				
35–39	DKTGF				
95–99	NSLTI				Chymotrypsinogen-A
29–33	DYETS				Lamprey hemoglobin

Single letter amino acid code has been used.

depending on the (ϕ, ψ) values of residues in the $(i+1)$st and $(i+2)$nd positions. The dihedral angles (ϕ_{i+1}, ψ_{i+1}), (ϕ_{i+2}, ψ_{i+2}) and (ϕ_{i+3}, ψ_{i+3}) given in Table 1 are quite different in many cases from those mentioned by Lewis *et al.* (1973). In particular, values of (ϕ_{i+3}, ψ_{i+3}) do not fall in any of the previously classified regions of 11 types of (ϕ, ψ) values. Thus the finding of these reversals is completely new in character. Model building studies made using Kendrew wire models and computer graphics have suggested that the stabilizing force for these loops is not only favorable non-bonded distances but also the main chain-side chain hydrogen bond and main chain-main chain hydrogen bond in adjacent peptide units, mainly of the five-membered type. Semi-empirical energy claculations are in progress to throw light on this.

In Table 2 the data on chain reversals obtained using the algorithm is presented in column (c). For comparison similar data obtained by Chou & Fasman (1977) and Rose & Seltzer (1977) are presented in columns (e) and (f) of this table. Columns (c), (e) and (f) of this table indicate that there is difference in the total number of chain reversals predicted by each of these methods. It should be mentioned here that the total number of bends obtained by our method include one turn each at the *N*- and *C*-terminals of each helix present in the protein. In other words, we have assumed that the helix is nothing but a large number of continuous turns or chain reversals. This assumption is supported by preliminary conformational energy calculations on tripeptides of alanine. These calculations have given energy minima at the $(-60°, -30°)$ and $(-60°, -30°)$ regions

TABLE 2

Analysis of amino acid residues in bends of different globular proteins

Protein	(a)	(b)	(c)	(d)	(e)	(f)
Lysozyme	129	84	27	20	20	19
Subtilisin BPN'	275	136	37	26	29	37
Hemoglobin α	141	71	24	14	–	–
Hemoglobin β	146	76	27	15	–	–
Flavodoxin	138	66	21	16	17	15
Rubredoxin	53	36	12	12	11	7
α-Chymotrypsin	245	125	37	37	32	36
Concanavalin A	237	101	28	28	28	34
Thermolysin	316	154	52	43	39	–
Chymotrypsinogen-A	245	111	33	28	–	–
HIPIP	85	67	22	19	17	16
Lactate dehydrogenase	331	145	43	32	33	44
Parvalbumin	108	70	22	9	–	–
Cytochrome b_5	85	52	15	7	6	10
BPTI	58	23	8	6	4	9
Carboxypeptidase A	307	140	44	34	36	39
Ribonuclease-S	124	54	15	9	11	22
Staphylococcal nuclease	142	75	23	16	12	–
Lamprey hemoglobin	148	79	27	8	7	–
Myoglobin	153	70	23	12	9	13
Cytochrome C_2	112	67	21	17	–	14
GPD	334	163	48	35	–	45

[a] Total number of residues (T_R).
[b] Total number of residues in bends (N_R)
[c] Number of bends from our calculations (T_B).
[d] Number of bends after deletion of one turn each from *N*- and *C*-terminals of helical region which are considered to be bend regions.
[e] Number of turns predicted by Chou & Fasman (1977).
[f] Number of turns predicted by Rose & Seltzer (1977).

(unpublished work, Prashanth & Kolaskar). In column (d) of this table, we have given the total number of bends obtained excluding the turns which formed the *N*- and *C*-terminals of the helices, to show that this column is in good agreement with column (e) obtained by Chou & Fasman (1977). This is quite understandable as the algorithms used in these two cases are quite similar.

It can be seen from columns (a) and (c) of Table 2 and Fig. 1(a) that there is a good correlation between the total number of residues in a given protein and the number of bends in that protein. This can be expressed in the relation

$$T_B = 0.122T_R + 5.94 \quad (1)$$

where T_B and T_R are, respectively, the total number of bends and amino acid residues in a particular protein. The correlation coefficient for least square line fit was found to be 0.95, indicating that the fit is of good quality. A similar relation was obtained by Rose & Wetlaufer (1977), using criteria developed to detect bends by Rose & Seltzer (1977). However, in that relation, proteins like myoglobin or hemoglobin do not follow the linearity. In the present analysis one does not have to exclude this type of protein.

To understand the reason for the linear relationship mentioned in eqn. 1, we have calculated the total number of hydrophobic and non-hydrophobic residues in each protein.

Our previous analysis (Kolaskar & Soman, 1979) on a large number of globular proteins has shown that the total number of non-hydrophobic residues (T_{NHB}) is linearly related to the total number of residues in that protein in the relation,

$$T_{NHB} = 0.619T_R - 2.164 \quad (2)$$

Also from our present analysis, we have found that the total number of bends (T_B) is linearly related to the total number of non-hydrophobic residues (T_{NHB}) in that protein in the relation which is shown in Fig. 1(b) and is given by

$$T_B = 0.204T_{NHB} + 6.03 \quad (3)$$

Thus, from eqn. 3 it becomes clear that the bends that generally occur on the surface of the protein (Kuntz, 1972) and non-hydrophobic residues, which are also on the exterior of the protein, are directly related. Thus eqn. 1 seems to be the result of eqns. 3 and 2. The thermodynamic stability of proteins and polypeptides requires eqn. 3. We have also noticed that the number of consecutive bends is not small and so the total number of residues in bends cannot be obtained simply by multiplying the number of bends by four (the number of residues in each bend). The analysis which we have carried out on the actual number of residues in bends in each protein (N_R) and the number of bends (T_B) indicates that there exists a linear relationship between these two quantities (see Fig. 2(a)). This can be expressed as

$$T_B = 0.293N_R + 1.5 \quad (4)$$

The correlation coefficient for least square fit was found to be 0.98. Thus it can be seen from columns (b) and (c) of Table 2 and eqn. 4 that approximately three independent amino acid residues form one turn. Eqns. 1 and 4 suggest that there must be a linear relation between

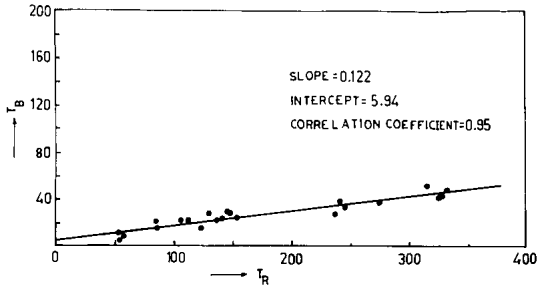

FIGURE 1(a)

Variation of total number of bends (T_B), as detected using present algorithm, against total number of amino acid residues (T_R) in each protein.

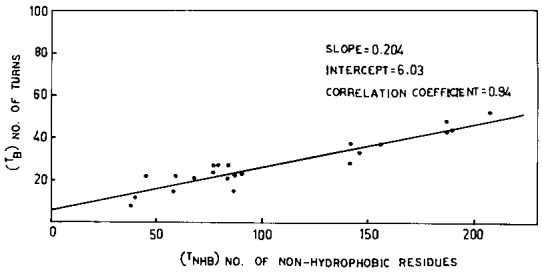

FIGURE 1(b)

Variation of total number of turns (T_B) versus total number of non-hydrophobic residues (T_{NHB}) in 21 globular proteins.

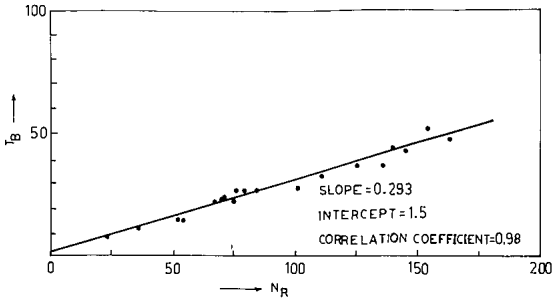

FIGURE 2(a)

Total number of bends (T_B) versus total number of residues in bends (N_R) in each protein.

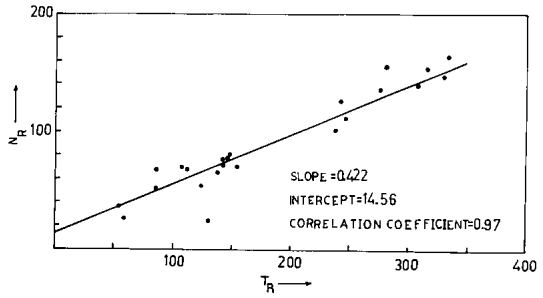

FIGURE 2(b)

Total number of residues in bends (N_R) versus total number of residues (T_R) in each protein. The slope of straight line indicates that more than 42% residues in each protein are in bends.

TABLE 3

Frequency of occurrence of amino acid in β-turns

Residues	i	i + 1	i + 2	i + 3	Total
Ala	39(10.37)	65(17.29)	36 (9.57)	52(13.83)	376
Val	32(10.49)	30 (9.84)	23 (7.54)	36(11.80)	305
Leu	34(11.56)	35(11.90)	32(10.86)	32(10.86)	294
Ile	24(12.57)	17 (8.90)	15 (7.85)	24(12.57)	191
Pro	28(20.29)	52(37.68)	9 (6.52)	23(16.67)	138
Trp	9(13.64)	3 (4.55)	12(18.18)	15(22.73)	66
Phe	25(18.25)	14(10.22)	28(20.44)	19(13.87)	137
Met					
Cys	13(17.57)	8(10.81)	12(16.22)	11(14.86)	74
Gly	47(14.42)	40(12.27)	72(22.09)	75(23.01)	326
Asn	50(27.78)	27(15.00)	42(23.33)	30(16.67)	180
Gln	13(11.40)	21(18.42)	13(11.40)	21(18.42)	114
Ser	73(22.12)	56(16.97)	52(15.76)	44(13.33)	330
Thr	46(18.70)	34(13.82)	39(15.85)	31(12.60)	246
Tyr	19(13.77)	27(19.57)	22(15.94)	27(19.57)	138
His	24(24.24)	13(13.13)	25(25.25)	12(12.12)	99
Lys	41(14.34)	53(18.53)	49(17.13)	46(16.08)	286
Arg	8 (7.69)	22(21.15)	13(12.50)	10 (9.62)	104
Asp	59(24.18)	43(17.62)	75(30.74)	44(18.03)	244
Glu	23(12.04)	37(19.37)	29(20.42)	37(19.37)	191

The total occurrence (n_t) of each residue in the positions 1st, 2nd, 3rd and 4th are represented by i, i + 1, i + 2, and i + 3. n is the total occurrence of each residue in 21 proteins. The frequency of occurrence, in percentage, of each residue in β-turn, $f_t = (n_t/n \times 100)$ is given in brackets.

total number of residues (T_R) and number of residues in bends (N_R) in a particular protein. This linear relation is shown in Fig. 2(b) and is expressed as

$$N_R = 0.422T_R + 14.56 \qquad (5)$$

Eqn. 5 indicates that in globular proteins more than 42% residues are in bends.

Since there are a large number of consecutive bends in proteins, it becomes essential to calculate the frequency of occurrence of each type of amino acid residue occupying a particular position – i to i + 3 in bends – rather than calculating only its probability of occurrence in bends irrespective of its position in bends. When one wants only the probability of a particular amino acid residue to occur in chain reversals, it is sufficient if one computes the number of times it occurs in bends irrespective of its position, so one should count that residue only once, even though it might be occurring in consecutive bends and occupying different positions, as has been done by Chou & Fasman (1977). But when one computes the frequency of occurrence of an amino acid residue in a particular position of β-bends one should count it as many times as it occurs in that position, which means that one will have to count it more than once when it occurs in consecutive bends. If Ala occurred as residue 12 in turns between 9 and 12 and 11 and 14, it should be counted twice, once in (i + 3)rd position of first turn and once in (i + 1)st position of second turn. The frequency of occurrence of each amino acid residue in (i)th, (i + 1)st, (i + 2)nd and (i + 3)rd positions is calculated using our data. These results are presented in Table 3. In most of the cases the trend of our results agrees with that of Chou & Fasman (1977), but the actual numbers obtained here are quite different.

It can be seen from Table 3 that only a very small percentage of Ile, Leu, and Val residues form part of bends and these residues do not have a preference for any particular position in bends, unlike Pro and Gly, which prefer (i + 1)st

TABLE 4

Proline occurring in the (i + 2)nd position of β-bends. In all cases the peptide unit between (i + 1)st and (i + 2)nd residue is found to be in the cis *conformation*

No.	Protein	Residue No.
1.	Subtilisin BPN'	168
2.	Papain	152
3.	Thermolysin	51
4.	Dehydrogenase	344
5.	Ribonuclease-S	93
6.	Ribonuclease-S	114
7.	Staphylococcal nuclease	47
8.	Staphylococcal nuclease	117
9.	Bence Jones REI	95
10.	Bence Jones REI	8
11.	Immunoglobulin G	451
12.	Carbonic anhydrase C	29
13.	Carbonic anhydrase C	200
14.	Carbonic anhydrase B	29
15.	Midge larva hemoglobin	74

and (i + 2)nd positions, respectively. Ala seems to be present in bends more frequently than Val, Leu, or Ile and prefers (i + 1)st position though not as strongly as Pro. Lys does not seem to have any positional preference, though a considerable number of times it has appeared in bends. Our analysis of proline residues in bends has shown that there are very few cases when Pro occurs in position (i + 2) of the bend and in all these cases the peptide bond between (i + 1)st and (i + 2)nd residue is found to be in the *cis* conformation. All such cases in which Pro occurs in the (i + 2)nd position are listed in Table 4. Secondly, we observed that almost all bends which contain Pro in the (i + 1)st position and are followed by amino acid residues other than Gly were only of the type having an average value of (ϕ_{i+1}, ψ_{i+1}) and (ϕ_{i+2}, ψ_{i+2}) as $(-50°, -30°)$, $(-100°, 0°)$ respectively. However, if Gly occurs in (i + 2)nd position then restrictions imparted because of side chains of L-amino acid residues in (i + 2)nd positions are not present and therefore a variety of bends are possible.

Thus there seems to be definite preference for a particular amino acid residue to occupy particular position in bends. It is also quite accepted that the tertiary structure or the folding of the protein is sequence-dependent. Thus even in bends one can expect certain preferences not only to individual amino acid residues but also to the pairs of these residues occurring in the (i + 1)st and (i + 2)nd positions. If this argument is valid, then the probability values calculated for a single amino acid residue to be in β-bends will have less significance because a β-bend-forming amino acid residue may not always form β-bends in association with certain types of amino acid residues. In a similar way, a β-bend breaker may turn out to have a very high potential to be in the bends when paired with certain amino acid residues. With this in mind, we have calculated the frequency of occurrence of all 400 pairs of amino acids in bends, using the procedure mentioned earlier. The results of this study are shown in Table 5.

From previous observations, Ala is a β-bend breaker. From Table 5 we can see that Ala prefers (i + 1)st position to (i + 2)nd position, as has been noted from Table 3. The additional and important information we get from Table 5 is that Ala-Asp and Ala-Gly pairs have a high probability to be in bends. The β-bend-forming capacity of Glu is very low. But when it is associated with Asp, Glu-Asp occurs frequently in bends (87.5%). From previous studies by others and also from our study, Gly is characterized as a strong β-bend former, preferring position (i + 2). When we consider the pairs Asn-Gly (57.14%), Gln-Gly (66.67%), Met-Gly (60%), Ala-Gly (42.86%), Arg-Gly (42.86%) and Ser-Gly (40.74%) they show a greater probability to occur in bends. However, the probability of occurrence of pair Pro-Gly (35.71%), is quite low when compared to other combinations mentioned above. It is known that Pro prefers (i + 1)st position and Gly (i + 2)nd position. But their combination, Pro-Gly, though in correct order of sequence, has less probability to occur in bends. Asp is another bend-forming amino acid residue which prefers to occur in position (i + 2), The pairs Ala-Asp (56.52%) Glu-Asp (87.51%), Lys-Asp (45.45%) and Pro-Asp (50%) have a high probability to occur in bends over other pairs. But the probability for Ser-Asp is quite low though Ser prefers (i + 1)st and Asp prefers (i + 2)nd position in bends.

TABLE 5(a)

Probability of occurrence of pairs in protein sequence considered

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	51	4	23	13	11	28	14	14	29	30	6	11	9	8	8	30	16	29	8	10
C	7	0	4	2	0	9	1	2	9	9	3	4	1	3	4	7	2	5	0	1
D	32	5	13	15	9	24	1	17	16	17	3	12	10	4	4	21	9	21	4	6
E	16	1	8	12	14	9	2	6	16	21	5	11	5	5	6	10	12	13	4	4
F	9	3	8	13	6	11	2	9	16	11	3	5	5	1	5	8	12	6	1	6
G	24	7	28	15	9	18	5	21	23	19	3	10	10	8	14	29	18	32	5	11
H	8	5	0	6	7	12	2	5	6	5	2	2	10	0	1	7	4	6	0	5
I	13	3	14	9	6	17	3	15	11	11	2	14	11	7	5	10	9	17	3	6
K	27	1	22	8	10	17	10	18	24	26	7	12	5	1	4	20	16	25	2	22
L	18	5	10	15	9	20	14	16	26	20	4	7	12	12	9	28	21	24	4	5
M	4	0	6	3	4	5	1	3	10	0	2	3	1	2	2	4	4	6	1	1
N	11	6	5	10	10	14	2	9	7	9	2	9	6	6	7	10	15	12	8	5
P	13	3	18	14	2	14	3	2	5	9	3	8	3	0	1	13	4	13	3	4
Q	13	0	9	3	2	9	6	5	10	5	21	6	3	6	5	5	7	6	1	2
R	7	3	3	3	3	7	2	5	3	14	3	7	2	4	0	12	4	9	2	5
S	30	5	17	15	12	27	7	10	20	19	6	10	5	14	7	31	33	20	9	14
T	21	6	14	9	13	20	5	8	15	19	2	11	16	9	4	19	19	17	4	9
V	36	12	24	11	7	20	10	12	18	24	3	16	9	9	6	29	16	27	3	11
W	5	2	3	2	1	9	1	4	3	2	1	1	2	3	3	4	7	9	1	1
Y	10	1	12	3	3	20	2	2	8	6	2	7	7	4	3	13	13	5	3	6

In forming the pairs, amino acid residue mentioned in vertical column will occupy the first position and residue mentioned along horizontal line will associate with it in position two.
Single letter amino acid code has been used.

TABLE 5(b)

Probability of occurrence of pairs in bends

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	10	2	13	1	0	12	2	2	5	4	0	4	2	2	0	6	3	2	2	1
C	1	0	0	0	0	3	0	0	0	2	0	0	0	0	0	2	1	0	0	0
D	6	0	3	6	4	9	0	1	3	0	0	5	0	0	2	3	1	5	0	1
E	4	0	7	4	2	3	0	0	3	4	1	3	0	3	0	4	2	0	1	1
F	0	0	0	2	1	2	0	0	3	2	0	0	0	0	0	3	0	0	0	2
G	0	1	7	2	4	3	2	0	4	2	1	0	1	1	0	5	4	6	1	3
H	1	3	0	0	1	3	1	0	2	2	0	0	0	0	0	1	0	0	0	2
I	1	1	4	1	1	2	1	2	2	1	0	1	0	0	0	0	0	1	0	0
K	2	1	10	1	1	6	4	0	5	5	0	5	1	0	0	2	5	3	1	7
L	2	2	3	4	2	2	1	1	4	2	0	2	2	0	2	2	5	1	0	3
M	0	0	2	0	0	3	1	0	3	0	0	0	0	0	0	1	0	0	0	0
N	0	0	1	3	2	8	1	1	0	1	0	2	1	1	0	0	6	0	0	0
P	4	1	9	10	1	5	1	0	5	2	0	5	0	0	0	4	2	1	2	0
Q	0	0	3	2	1	6	3	1	0	2	0	1	0	0	0	0	2	0	0	1
R	0	2	1	0	1	3	2	2	1	4	0	1	0	2	0	1	0	1	2	1
S	0	1	6	3	4	11	2	1	5	2	0	4	0	3	4	6	8	1	2	1
T	2	0	3	1	2	5	2	2	6	0	0	3	0	1	1	7	0	4	1	2
V	5	0	4	1	0	1	3	1	1	2	0	6	0	2	1	3	2	1	0	1
W	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	1	0
Y	3	1	4	2	2	7	0	1	1	1	0	0	2	0	1	4	1	0	2	0

TABLE 5(c)

Frequency of occurrence of pairs in bends (in percentage)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	20	50	57	7	0	43	14	14	17	13	0	36	22	25	0	20	19	7	25	10
C	14	–	0	0*	–	33	0	0*	0	22	0*	0	0*	0*	0	29	50*	0	–	0*
D	19	0	23	40	44	38	0*	6	19	0	0*	42	0	0	50*	14	11	24	0	17
E	25	0	88	33	14	33	0*	0	19	19	20	27	0	60	0	40	17	0	25	25
F	0	0*	0	15	17	18	0*	0	19	18	0*	0	0	0*	0	38	0	0	0*	33
G	0	14	25	13	44	17	40	0	17	11	33	0	10	13	0	17	22	19	20	27
H	13	60	–	0	14	25	50*	0	33	40	0*	0*	0	–	0*	14	0	0	–	40
I	8	33	29	11	17	12	33	13	18	9	0*	7	0	0	0	0	0	6	0*	0
K	7	100*	45	13	10	35	40	0	21	19	0	42	20	0	0	10	31	12	50*	32
L	11	40	30	27	22	10	7	6	15	10	0	29	17	0	22	7	24	4	0	60
M	0	–	33	0*	0	60	100*	0*	30	–	0*	0*	0*	0*	0*	25	0	0	0*	0*
N	0	0	20	30	20	57	50*	11	0	22	0*	22	17	17	0	0	40	0	0	0
P	31	33	50	71	50	38	33	0*	100	22	0*	63	0*	–	0*	31	50	8	67*	0
Q	0	–	33	67*	50*	67	50	20	0	40	0	17	0*	0	0	0	29	0	0*	50
R	0	67*	33	0*	33	43	100*	40	33	29	0*	14	0*	50	–	8	0	11	100*	20
S	0	20	35	20	33	41	29	10	25	11	0	40	0	21	57	19	24	5	22	7
T	10	0	21	11	15	25	40	25	40	0	0*	27	0	11	25	37	0	24	25	22
V	14	0	17	9	0	5	30	8	6	8	0*	38	0	22	17	10	13	4	0*	9
W	0	0*	0*	0*	0*	0	0	25	0*	0*	0*	0*	0*	0*	0*	50	0	0	100*	0*
Y	30	100*	33	67*	67*	35	0*	50*	13	17	0*	0	29	0	33	31	8	0	67*	0

0 indicates that that particular pair had zero probability to occur in bends though it had certain probability to occur in protein sequences considered.

– indicates that that particular pair had zero probability to occur in protein sequence considered.

* indicates that the pair occurs in protein sequences considered three or less times.

Proline, another bend-forming residue, occurring in position (i + 1) in association with lysine (Pro-Lys) (100%), has a unit probability to occur in bends. Notable associations are, Pro-Glu (71.43%), Pro-Asn (62.5%), Pro-Thr (50%) and Pro-Trp (66.67%).

When an amino acid residue associates with itself, the probability of these pairs to occur in bends is zero for Cys, Met, Pro, Gln, Arg, Thr and Tyr. Even for other amino acids this probability is not significant. Arg is normally assumed to be β-bend breaker (Levitt, 1978). As can be seen from Table 5 the pairs Arg-Cys (61.67%), Arg-His (100%) and Arg-Trp (100%) have frequency values which indicate that Arg is more prone to be in bends as the said pairs.

Thus the contents and discussion of Table 5 clearly suggest that instead of categorizing a single amino acid residue to be β-bend former, indifferent to, or breaker, one should look at the probability of occurrence of each pair of amino acid residues in β-bends in order to predict them in proteins. This becomes essential mainly because chain reversals consist of only four to five residues as against 10 or 12 residues in helix or β-strands. Therefore to predict this type of secondary structure in proteins, the consideration of probability values of single amino acid residues alone may not be sufficient.

REFERENCES

Chou, P.Y. & Fasman, G.D. (1977) *J. Mol. Biol.* **115**, 135–175

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

Levitt, M. (1978) *Biochemistry (Wash.)* **17**, 4277–4285

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

Kolaskar, A.S. & Soman, K.V. (1979) *Curr. Sci.* **48**, 928–931

Kuntz, I.D. (1972) *J. Am. Chem. Soc.* **94**, 4009–4012

Rose, G.D. (1978) *Nature* **272**, 586–590

Rose, G.D. & Seltzer, J.P. (1977) *J. Mol. Biol.* **113**, 153–169

Rose, G.D. & Wetlaufer, D.B. (1977) *Nature* **268**, 769–770

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

Address:

Dr. *A.S. Kolaskar*
School of Life Sciences
University of Hyderabad
"Golden Threshold" Nampally
Station Road
Hyderabad - 500 001
India